SIAM J. Applied Dynamical Systems
Vol. 3, No. 1, pp. 1–68

Defects in Oscillatory Media: Toward a Classification[*]

Björn Sandstede[†] and Arnd Scheel[‡]

Abstract. We investigate, in a systematic fashion, coherent structures, or defects, which serve as interfaces between wave trains with possibly different wavenumbers in reaction-diffusion systems. We propose a classification of defects into four different defect classes which have all been observed experimentally. The characteristic distinguishing these classes is the sign of the group velocities of the wave trains to either side of the defect, measured relative to the speed of the defect. Using a spatial-dynamics description in which defects correspond to homoclinic and heteroclinic connections of an ill-posed pseudoelliptic equation, we then relate robustness properties of defects to their spectral stability properties. Last, we illustrate that all four types of defects occur in the one-dimensional cubic-quintic Ginzburg–Landau equation as a perturbation of the phase-slip vortex.

Key words. pattern formation, coherent structures, spatial dynamics, group velocity

AMS subject classifications. 37L10, 35K57, 34C37

DOI. 10.1137/030600192

1. Introduction.

In this paper, we investigate coherent structures in essentially one-dimensional spatially extended systems. Specifically, we are interested in interfaces between stable spatially periodic structures with possibly different spatial wavenumbers as illustrated in Figure 1.1. These interfaces can also be thought of as defects at which the underlying perfectly periodic structure is broken. In many cases, both the periodic structures and the defect will depend on time. We focus on defects where the resulting pattern is time-periodic, possibly after transforming into an appropriate moving frame of reference. Our goal is to investigate the existence and stability properties of such defects. In particular, we are interested in classifying defects according to their codimension and studying their robustness under parameter variations. Throughout this paper, we will use the term *wave trains* to denote spatially periodic travelling waves.

We begin by briefly reviewing some numerical simulations and experiments in which defects have been observed and by introducing, on a heuristic level, the concepts needed for the classification of defects. Afterward, we recall some facts we need about wave trains before stating the definition of defects and our main results. Table 1.1 contains a summary of the notation we shall use throughout this paper.

[*] Received by the editors May 6, 2003; accepted for publication (in revised form) by M. Golubitsky October 21, 2003; published electronically February 24, 2004.

http://www.siam.org/journals/siads/3-1/60019.html

[†] Department of Mathematics, The Ohio State University, Columbus, OH 43210 (sandstede.1@osu.edu). The research of this author was partially supported by the Alfred P. Sloan Foundation and by the National Science Foundation through grant DMS-0203854.

[‡] Department of Mathematics, University of Minnesota, Minneapolis, MN 55455 (scheel@math.umn.edu). The research of this author was partially supported by the National Science Foundation through grant DMS-0203301. This author is grateful to the Sonderforschungsbereich 555 and the Fachbereich Mathematik at the FU Berlin for their kind support during a one-month research visit.

Table 1.1
Notation.

$u(x,t)$	solution to reaction-diffusion system
$u_{\mathrm{wt}}(kx-\omega t;k)$	wave train (2π-periodic in argument ϕ)
$\phi = kx - \omega t$	travelling-wave coordinate (wave train)
k	wavenumber
ω	temporal frequency
$\omega_{\mathrm{nl}}(k)$	nonlinear dispersion relation
$c_{\mathrm{p}} = \omega_{\mathrm{nl}}(k)/k$	phase velocity
$c_{\mathrm{g}} = \mathrm{d}\omega_{\mathrm{nl}}(k)/\mathrm{d}k$	group velocity
$\check{\lambda}$, $\lambda_{\mathrm{lin}}(\nu)$	wave-train eigenvalue and linear dispersion relation computed in frame moving with speed c_{p}
λ	temporal Floquet exponent
$\rho = \exp(2\pi\lambda/\omega_{\mathrm{d}})$	temporal Floquet multiplier
ν	complex spatial Floquet exponent
γ	spatial Floquet exponent
c_{d}	speed of defect
ω_{d}	temporal frequency of defect
$\xi = x - c_{\mathrm{d}}t$	travelling-wave coordinate (defect)
$\tau = \omega_{\mathrm{d}}t$	rescaled time (2π-periodic)
$u_{\mathrm{d}}(\xi,\tau)$	defect (2π-periodic in τ)
1	identity operator
$\mathrm{N}(\mathcal{L})$, $\mathrm{Rg}(\mathcal{L})$	null space and range of a closed linear operator $\mathcal{L}$
$i(\mathcal{L}) = \dim \mathrm{N}(\mathcal{L}) - \operatorname{codim} \mathrm{Rg}(\mathcal{L})$	index of a Fredholm operator $\mathcal{L}$
$Y = H^{1/2}(S^1,\mathbb{R}^n) \times L^2(S^1,\mathbb{R}^n)$	spaces for spatial-dynamical systems
$Y^1 = H^1(S^1,\mathbb{R}^n) \times H^{1/2}(S^1,\mathbb{R}^n)$	

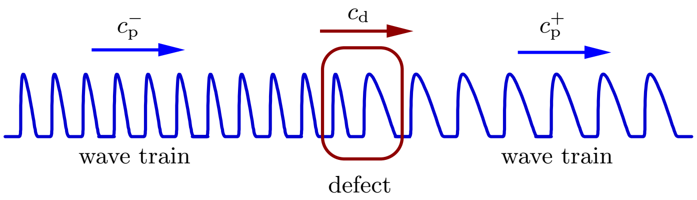
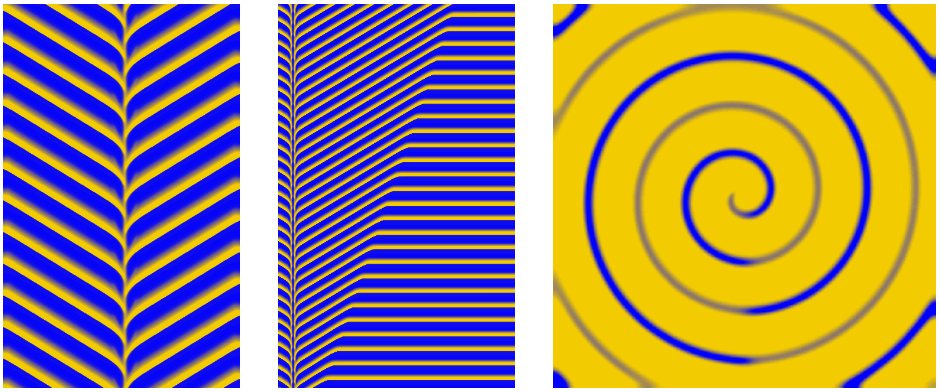


Figure 1.1. *A defect travelling with speed c_{d} through spatially periodic structures that themselves travel with phase velocities c_{p}^- behind and c_{p}^+ ahead of the defect.*

1.1. Motivation.

Experiments and simulations. To set the scene, we describe numerical simulations of defects and review various experiments in which they have been observed. Consider first the left and center plot in Figure 1.2. Both are space-time contour plots of solutions to the Brusselator (see Appendix B for the equations). Figure 1.2(i), which reproduces simulations from [38], shows a standing defect that emits wave trains alternately to the left and right so that the emitted wave trains travel away from the defect toward the domain boundary. Defects of this type are often referred to as flip-flops or one-dimensional spirals. Note that the defect is time-periodic since the space-time plot is periodic in the vertical time direction. Figure 1.2(ii) shows a standing defect near the left domain boundary that emits wave trains simultaneously to the left and right; such defects are often referred to as target patterns. In the interior of the domain, a travelling defect is formed between the waves emitted by the target



(i) Flip-flop [MOVIE] (ii) Target and sink [MOVIE] (iii) Line defect [MOVIE]

Figure 1.2. *The numerical simulations shown here were carried out using Barkley's code* EZSPIRAL [3]. *Figure* (i) *on the left shows a space-time plot (time is plotted upward and space horizontally) of a flip-flop that emits wave trains alternately to the left and right: The emitted wave trains travel away from the defect toward the domain boundary. Figure* (ii) *in the middle is a space-time plot (time is plotted upward and space horizontally): Near the left boundary, we see a standing target pattern that emits waves simultaneously to the left and right. In the interior of the domain, a travelling defect is formed between the waves emitted by the target pattern and the spatially homogeneous oscillations that occupy the right half of the domain. Figure* (iii) *is a snapshot of a two-dimensional spiral wave. A one-dimensional line defect emerges from the center of the spiral and connects to the bottom of the domain. Along the line defect, the phase of the oscillations jumps by half a period.*

pattern and the spatially homogeneous oscillations that occupy the right half of the domain. The travelling defect is time-periodic when viewed in a comoving frame. Indeed, shearing the space-time plot appropriately in the horizontal direction renders the figure vertically periodic with the defect being a vertical line.

Flip-flops of a slightly different nature have been observed in numerical simulations in the excitable regime of the Oregonator [40], where a localized pulse destabilizes and releases pulses in its wake—a mechanism often referred to as backfiring. The emerging pattern resembles Figure 1.2(i) with pulses being released alternately to the left and right. The resulting interfaces can also move with nonzero speed [40].

Chemical oscillations have been generated in various reactions, most of which are related to the original Belousov–Zhabotinsky mechanism. In [38], defect patterns were observed in the chlorite-iodite-malonic-acid (CIMA) reaction in the parameter regime where stable, stationary, spatially periodic Turing patterns and time-periodic spatially homogeneous oscillations coexist. In one space dimension, [38, Figure 2] shows flip-flops of the type plotted in Figure 1.2(i). We emphasize that the waves emerging from the chemical flip-flop in [38] are not generated by an inhomogeneity in the medium. Instead, the defect forms spontaneously. At the defect, the phase of the wave trains jumps by half a period.

Defects with a different type of phase slip were observed in photo-sensitive monolayers on thin Belousov–Zhabotinsky reaction solutions [58]. The two-dimensional spiral waves shown in [58, Figures 2 and 5] exhibit stationary line defects along which the phase of the oscillations jumps by half the period. The spiral waves are presumably generated by a period-doubling bifurcation of spatially homogeneous oscillations that then leads to a Hopf bifurcation of the

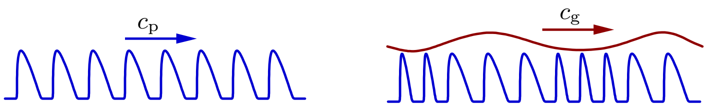


Figure 1.3. *The difference between the phase velocity c_{p} and the group velocity c_{g} of wave trains. The latter describes the speed with which slowly varying modulations of the wavenumber propagate.*

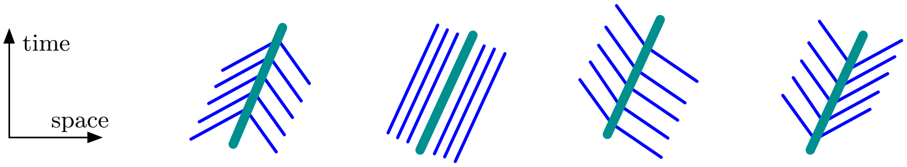


Figure 1.4. *A sketch of the characteristic curves that enter or leave each defect depending on the speed of the group velocities of the wave trains to the left and right of the defect compared with the speed of the defect. From left to right: Sinks, contact defects, transmission defects, and sources.*

two-dimensional spirals [52]. The line defects appear to orient themselves parallel to the propagation direction of the wave trains. We refer to Figure 1.2(iii) for a snapshot of a two-dimensional spiral wave that exhibits such a line defect. The pattern shown there was first found in [18], to which we refer for details on the equation used to generate it.

Hydrothermal waves can also exhibit defects. The recent experiments in [1, 36, 37], in which nonlinear waves and various kinds of defects were triggered by heated wires immersed in thin layers of oil, were motivated by the desire to obtain a quantitative comparison with coupled Ginzburg–Landau equations. Other experiments where defects have been observed are the printer instability [19], laterally heated fluid layers [4], and thermal convection of binary fluids [31].

Heuristic classification. We are interested in finding characteristic properties of coherent structures of the kind shown in Figure 1.1. It turns out that the group velocities of the asymptotic wave trains are the deciding characteristic. The group velocity c_{g} associated with a wave train can be thought of as the speed with which small perturbations are transported along the wave train (we will make this more precise in section 1.3 below; see also Figure 1.3). Alternatively, the group velocity can be computed as the derivative of the frequency of the wave train with respect to its wavenumber (see section 1.2 below). We may then distinguish defects according to whether perturbations to the left or right of the defect travel toward, parallel to, or away from the interface. In fact, we propose the following classification:

sinks	$c_{\mathrm{g}}^- > c_{\mathrm{d}} > c_{\mathrm{g}}^+$,
contact defects	$c_{\mathrm{g}}^- = c_{\mathrm{d}} = c_{\mathrm{g}}^+$,
transmission defects	either $c_{\mathrm{g}}^\pm > c_{\mathrm{d}}$ or $c_{\mathrm{g}}^\pm < c_{\mathrm{d}}$,
sources	$c_{\mathrm{g}}^- < c_{\mathrm{d}} < c_{\mathrm{g}}^+$,

where c_{d} is the speed of the defect, and c_{g}^- and c_{g}^+ denote the group velocities of the wave trains to the left and right, respectively, of the interface. We refer to Figure 1.4 for an illustration. We emphasize that the slope of the level sets in the space-time plots of Figure 1.2 reflects the *phase* velocity of the wave trains. In contrast, the characteristic curves sketched in Figure 1.4

indicate the direction of the *group* velocity. In general, the signs of these two velocities are not related. While phase and group velocity of the waves in Figure 1.2(i) and (ii) have the same sign, the group velocity of the waves in Figure 1.2(iii) is directed toward the boundary, while they travel toward the spiral center.

The intuition is that sources generate the wave trains to either side since their group velocity points away from the interface. Sinks are passively created by wave trains that transport from the left and right toward the interface, thus forming it. Contact and transmission defects typically connect identical wave trains; these defects account for phase differences between the wave trains to their left and right.

We remark that there are defects that do not fit into the classification shown above (for instance, those for which $c_{\mathrm{g}}^{-} = c_{\mathrm{d}} > c_{\mathrm{g}}^{+}$). Nevertheless, we will argue in section 6.10 that the only "relevant" defects are those captured by our classification.

Last, we briefly revisit the experiments and simulations mentioned above to demonstrate that each of the defect classes indeed arises in physical systems. The flip-flop in Figure 1.2(i), the target pattern in Figure 1.2(ii), and the chemical flip-flops observed by [38] in the CIMA reaction appear to be sources. The interface between travelling and standing waves shown in the center of Figure 1.2(ii) is a sink. We believe that the line defects that are illustrated in Figure 1.2(iii) and that occur in the modified Belousov–Zhabotinsky experiment [58] are contact defects. Transmission defects arise in the ferroin-catalyzed Belousov–Zhabotinsky reaction [20] and in the heated-wire experiments [1, 36]. Sources and sinks have also been observed in the printer instability [19] and in the heated-wire experiments [1, 36].

Defects in general reaction-diffusion systems. The idea to characterize coherent structures[1] using the group velocity is, of course, not new. Interfaces between wave trains with almost identical wavenumbers were, for instance, studied in great detail in [26]. Defects in the cubic-quintic and in coupled complex Ginzburg–Landau equations were investigated by van Saarloos and coworkers [41, 23, 22] and by Doelman [10].

Our goal is to investigate defects arising in reaction-diffusion equations. The difficulty is that defects cannot be obtained as solutions to an ordinary differential equation (ODE). Instead, they are solutions to the partial differential equation (PDE) that depend genuinely on time. In fact, while the formation of periodic structures has been studied comprehensively in one, two, and three space dimensions (see [8] and the references therein), defects within these periodic structures are not as well understood, at least from a mathematical viewpoint. There are two reasons for this. First, defects are modulated waves and therefore time-periodic in a comoving frame. Thus, dynamical-systems methods with respect to the spatial variable, that are so successful when dealing with travelling waves, are not immediately applicable. Second, it appears difficult to use period maps to investigate defects as the linearization about a defect has the essential spectrum up to the imaginary axis. This fact precludes the immediate use of implicit function theorems. Note that the essential spectrum of each defect is generated by the spectrum of the asymptotic wave trains that always touches the imaginary axis. When studying spatially periodic patterns that respect a lattice symmetry group, this issue can be resolved by restricting our attention to functions that respect the same lattice group. Defects, however, inherently break the lattice symmetry.

[1] Throughout this paper, we shall use the terms *defect* and *coherent structure* interchangeably.

We focus on essentially one-dimensional media such as the real line $x \in \mathbb{R}$, cylindrical domains $\mathbb{R} \times \Omega \subset \mathbb{R} \times \mathbb{R}^m$, or patterns with a radial symmetry $x \in \mathbb{R}^+$. In these cases, we can reverse the role of time and space and treat the unbounded spatial variable x as the evolution variable, while time is restricted to the compact interval of periodicity. Dynamical-systems techniques can then be adapted to investigate bifurcations of small-amplitude solutions [30, 27, 55]. It is also possible to study periodic [34] as well as homoclinic and heteroclinic solutions [39, 49] of not necessarily small amplitude. Based on these ideas, we characterize in this paper properties of typical defects of arbitrary amplitude by interpreting them as homoclinic and heteroclinic trajectories that are constructed in a robust transverse fashion. To classify defects, we count relative dimensions of the infinite-dimensional stable and unstable manifolds.

Last, we do not wish to give the impression that all interesting patterns observed in nature are necessarily time-periodic. In fact, there are many fascinating patterns that do not fit at all into the framework described above and that are therefore not captured by the analysis presented here. However, as documented above, many patterns observed in physical systems are time-periodic when viewed in an appropriate frame and therefore fit into our framework.

1.2. Main results.

Reaction-diffusion systems. As a prototype for equations that give far-from-equilibrium dynamics, we focus on reaction-diffusion equations

$$u_t = Du_{xx} + f(u), \qquad x \in \mathbb{R}, \tag{1.1}$$

where $u \in \mathbb{R}^n$. We assume that the diffusion matrix is diagonal with strictly positive entries and that the nonlinearity $f \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$ is smooth. We think of f as a generic nonlinearity with no additional symmetries.

Wave trains. The common feature of the experiments mentioned above is the presence of wave trains which are travelling waves of the form $u_{\mathrm{wt}}(kx - \omega t; k)$, where $u_{\mathrm{wt}}(\phi; k)$ is 2π-periodic in ϕ. Typically, the spatial wavenumber k and the temporal frequency ω are related via the nonlinear dispersion relation $\omega = \omega_{\mathrm{nl}}(k)$ so that the phase velocity is given by $c_{\mathrm{p}} = \omega_{\mathrm{nl}}(k)/k$. A second quantity related to the nonlinear dispersion relation is the group velocity

$$c_{\mathrm{g}} = \frac{\mathrm{d}\omega_{\mathrm{nl}}}{\mathrm{d}k} \tag{1.2}$$

of the wave train which will play a central role in our results. The group velocity c_{g} gives the speed of propagation of small localized wave-package perturbations of the wave train (see Example II in section 1.3). Our primary interest is in wave trains that have a nonconstant dispersion relation[2] so that $c_{\mathrm{g}} \not\equiv 0$.

We shall assume that the wave trains are spectrally stable. If we linearize (1.1) about a wave train in the frame $\phi = kx - \omega_{\mathrm{nl}}(k)t$, we obtain the linear operator[3]

$$\mathcal{L}_{\mathrm{wt}} = k^2 D\partial_{\phi\phi} + \omega_{\mathrm{nl}}(k)\partial_\phi + f'(u_{\mathrm{wt}}(\phi; k)). \tag{1.3}$$

[2] Reversible Turing patterns, for instance, have a degenerate dispersion relation $\omega_{\mathrm{nl}}(k) \equiv 0$ so that the group velocity vanishes for all k. In this case, our proposed classification into four defect types does not make sense.

[3] Some of our expressions below are only valid for $k \neq 0$. The results, however, are also true when $k = 0$, and we will comment on this in section 3.3.

Spectral stability means that the spectrum of $\mathcal{L}_{\mathrm{wt}}$ on $L^2(\mathbb{R},\mathbb{C}^n)$ is contained strictly in the left half-plane except for a unique curve

$$\lambda_{\mathrm{lin}}(\nu) = a\nu + d\nu^2 + \mathrm{O}(\nu^3), \qquad \nu \in \mathrm{i}\mathbb{R}, \tag{1.4}$$

that touches the origin with a quadratic tangency $d > 0$. In this case, we actually know that $a = c_{\mathrm{p}} - c_{\mathrm{g}}$. We refer to section 3.1 for details.

Defects. Next, we consider coherent structures which are waves that are time-periodic in an appropriate frame of reference and asymptotic in space to wave trains with possibly different wavenumbers.

Definition 1.1. *We say that a solution $u(x,t) = u_{\mathrm{d}}(x - c_{\mathrm{d}}t, \omega_{\mathrm{d}}t)$ of (1.1) is an* elementary defect *with speed c_{d} and frequency ω_{d} if $\omega_{\mathrm{d}} \neq 0$ and if there are asymptotic wavenumbers k_- and k_+ and smooth phase-correction functions $\theta_\pm(\xi)$ with $\theta'_\pm(\xi) \to 0$ as $\xi \to \pm\infty$ such that*

$$u_{\mathrm{d}}(\xi,\tau) = u_{\mathrm{d}}(\xi,\tau + 2\pi) \tag{1.5}$$

and

$$u_{\mathrm{d}}(\xi,\omega_{\mathrm{d}}t) - u_{\mathrm{wt}}(k_\pm\xi + (k_\pm c_{\mathrm{d}} - \omega_{\mathrm{nl}}(k_\pm))t - \theta_\pm(\xi); k_\pm) \longrightarrow 0 \tag{1.6}$$

uniformly in t as $\xi \to \pm\infty$ for the functions as well as their derivatives with respect to (ξ, t). Last, we assume that $\partial_\xi u_{\mathrm{d}}(\xi,\tau)$ and $\partial_\tau u_{\mathrm{d}}(\xi,\tau)$ are linearly independent functions.

Formulated in the steady frame, the periodicity (1.5) and convergence (1.6) conditions are

$$u(x,t) = u(x + c_{\mathrm{d}}T_{\mathrm{d}}, t + T_{\mathrm{d}}), \qquad T_{\mathrm{d}} = \frac{2\pi}{\omega_{\mathrm{d}}},$$

and

$$u(x,t) - u_{\mathrm{wt}}(k_\pm x - \omega_{\mathrm{nl}}(k_\pm)t - \theta_\pm(x - c_{\mathrm{d}}t); k_\pm) \longrightarrow 0$$

uniformly in $0 \le t \le T_{\mathrm{d}}$ as $x \to \pm\infty$.

We will see in Corollary 5.2 that the phase functions $\theta_\pm(\xi)$ can be chosen to be constants $\theta_\pm$ for sinks, sources, and transmission defects, whereas contact defects have phase functions $\theta_\pm(\xi) \propto \log\xi + \mathrm{O}(1/|\xi|)$ that diverge logarithmically.

Throughout this paper, we denote the phase and group velocities of the asymptotic wave trains by

$$c_{\mathrm{p}}^\pm = \frac{\omega_{\mathrm{nl}}(k_\pm)}{k_\pm}, \qquad c_{\mathrm{g}}^\pm = \frac{\mathrm{d}\omega_{\mathrm{nl}}}{\mathrm{d}k}(k_\pm),$$

respectively. Note that (1.5)–(1.6) imply that

$$\omega_{\mathrm{nl}}(k_+) - k_+c_{\mathrm{d}} = \omega_{\mathrm{d}} = \omega_{\mathrm{nl}}(k_-) - k_-c_{\mathrm{d}}. \tag{1.7}$$

In particular, the assumption that defects are time-periodic implies the Rankine–Hugoniot condition

$$c_{\mathrm{d}} = \frac{\omega_{\mathrm{nl}}(k_+) - \omega_{\mathrm{nl}}(k_-)}{k_+ - k_-} \tag{1.8}$$

for the defect speed whenever $k_+ \neq k_-$. To justify our use of the term Rankine–Hugoniot, we note that the group velocity, which measures transport, is the derivative of the frequency. In conservation laws, transport is the derivative of the flux, so we may interpret the frequency ω_{nl} as the flux function of a fictitious conservation law. In that sense, (1.8) is the corresponding Rankine–Hugoniot condition. Using these findings, we see that (1.6) is equivalent to

$$u_{\mathrm{d}}(\xi,\tau) - u_{\mathrm{wt}}(k_\pm \xi - \tau - \theta_\pm(\xi); k_\pm) \longrightarrow 0, \qquad \xi \to \pm\infty. \tag{1.9}$$

To illustrate (1.8), consider the sink in the center of Figure 1.2(ii). Since the wave trains to the right of the defect are spatially homogeneous, we have $k_+ = 0$. Inspecting the figure further, we see that $\omega_{\mathrm{nl}}(k_-) > \omega_{\mathrm{nl}}(k_+)$, and (1.8) implies that $c_{\mathrm{d}} > 0$, which is consistent with the simulation.

Main result. We are now ready to describe the main result of this paper. We begin by introducing four distinct classes of defects. Each type occurs in an open set of reaction-diffusion systems, or, more precisely, for nonempty open subsets $\mathcal{U}$ of nonlinearities f in $C^3(\mathbb{R}^n,\mathbb{R}^n)$ and of diffusion matrices D. We define the four different defect types as follows:

(i) *Sinks* are elementary defects with $c_{\mathrm{g}}^- > c_{\mathrm{d}} > c_{\mathrm{g}}^+$.

(ii) *Contact defects* are elementary defects with $c_{\mathrm{g}}^- = c_{\mathrm{d}} = c_{\mathrm{g}}^+$.

(iii) *Transmission defects* are elementary defects with either $c_{\mathrm{g}}^\pm > c_{\mathrm{d}}$ or $c_{\mathrm{g}}^\pm < c_{\mathrm{d}}$.

(iv) *Sources* are elementary defects with $c_{\mathrm{g}}^- < c_{\mathrm{d}} < c_{\mathrm{g}}^+$.

All contact and transmission defects that we are aware of have $k_- = k_+$. In fact, contact defects for which $k_- \neq k_+$ and $\omega''(k_-) \neq \omega''(k_+)$ will not persist upon varying (k_-, k_+), and we will therefore restrict our analysis to contact defects for which $k_- = k_+$. Note also that we include neither sinks that have $k_- = k_+$ nor degenerate sinks for which $c_{\mathrm{g}}^- = c_{\mathrm{d}} > c_{\mathrm{g}}^+$ or $c_{\mathrm{g}}^- > c_{\mathrm{d}} = c_{\mathrm{g}}^+$, since we do not expect that such defects occur for open sets of wavenumbers k_- or k_+, respectively. We refer to section 6.10 for more details.

Our main result will link robustness properties of defects to spectral properties of the linearization of the period map of (1.1). To describe these properties, we therefore linearize (1.1) in the comoving frame $\xi = x - c_{\mathrm{d}}t$ with $\tau = \omega_{\mathrm{d}}t$ about an elementary defect $u_{\mathrm{d}}(\xi,\tau)$ and obtain the linear equation

$$\omega_{\mathrm{d}} u_\tau = D u_{\xi\xi} + c_{\mathrm{d}} u_\xi + f'(u_{\mathrm{d}}(\xi,\tau))u. \tag{1.10}$$

We denote the linear period map of this parabolic equation with time-periodic coefficients by

$$\Phi_{\mathrm{d}}:\ u(\cdot,0) \longmapsto u(\cdot,2\pi).$$

For any pair of real numbers $\eta = (\eta_-,\eta_+)$, we define $L^2_\eta(\mathbb{R},\mathbb{R}^n)$ to be the space of all locally square-integrable functions for which

$$\|u\|^2_{L^2_\eta} = \int_{\mathbb{R}_-} \left| u(\xi)\mathrm{e}^{\eta_- \xi} \right|^2 \mathrm{d}\xi + \int_{\mathbb{R}_+} \left| u(\xi)\mathrm{e}^{\eta_+ \xi} \right|^2 \mathrm{d}\xi$$

is finite.

Definition 1.2. *Consider Φ_{d} on $L^2_\eta(\mathbb{R},\mathbb{R}^n)$ with weights $\eta_\pm$ that are sufficiently close to zero and satisfy*

$$\operatorname{sign}\eta_\pm = \operatorname{sign}\left(c_{\mathrm{d}} - c_{\mathrm{g}}^\pm\right). \tag{1.11}$$

We say that an elementary defect is transverse *if it has* minimal spectrum *in the following sense:*

(i) *For* sinks*, we assume that $\Phi_\mathrm{d} - 1$ has a bounded inverse on $L^2_\eta(\mathbb{R},\mathbb{R}^n)$.*

(ii) *For* contact defects*, we assume that $\Phi_\mathrm{d} - 1$ has a bounded inverse $L^2_\eta(\mathbb{R},\mathbb{R}^n)$ for all weights $\eta_- = \eta_+ \neq 0$ that are sufficiently close to zero.*

(iii) *For* transmission defects*, we assume that $\rho = 1$ has algebraic multiplicity one as an eigenvalue of Φ_d on $L^2_\eta(\mathbb{R},\mathbb{R}^n)$.*

(iv) *For* sources*, we assume that $\rho = 1$ has algebraic multiplicity two as an eigenvalue of Φ_d on $L^2_\eta(\mathbb{R},\mathbb{R}^n)$.*

It turns out that, in each of the above four cases, $\Phi_\mathrm{d} - 1$ is Fredholm with index zero on $L^2_\eta(\mathbb{R},\mathbb{R}^n)$ with weights chosen according to Definition 1.2. The above requirements on the spectra of Φ_d can also be formulated in terms of multiplicities of appropriate roots of certain Evans functions (see section 5). We emphasize, however, that the minimal-spectrum assumption for contact defects does not refer to the usual Evans function that we constructed in [50] since the essential spectrum of contact defects always extends into the right half-plane in the exponentially weighted spaces that we use (see section 6.1).

We recall the main hypothesis that we require for the wave trains.

Hypothesis 1.3. *Each of the wave trains that we consider is part of a one-parameter family given locally by solutions $u_\mathrm{wt}(kx - \omega_\mathrm{nl}(k)t; k)$ of (1.1), where $u_\mathrm{wt}(\phi; k)$ is 2π-periodic in ϕ. We also assume that the dispersion relation $\omega = \omega_\mathrm{nl}(k)$ is well defined and smooth, again locally near each of the wave trains, and that $\omega''_\mathrm{nl}(k) \neq 0$. Last, we assume that each of these wave trains is spectrally stable in the sense sketched in (1.4) and made precise in Hypotheses 3.1 and 3.2 below.*

The following theorem distills the analysis of the present paper into a multiplicity and robustness result.

Theorem 1.4. *Assume that Hypothesis 1.3 is met. We then have the following:*

(i) Transverse sinks *occur in two-parameter families that are parametrized by the asymptotic wavenumbers $k_\pm$.*

(ii) Transverse contact defects *appear as one-parameter families that are parametrized by the asymptotic wavenumber $k_- = k_+$.*

(iii) Transverse transmission defects *appear as one-parameter families that are parametrized by the asymptotic wavenumber k_+ if $c_\mathrm{g}^\pm < c_\mathrm{d}$ (and by k_- if $c_\mathrm{g}^\pm > c_\mathrm{d}$).*

(iv) Transverse sources *appear for a discrete set of wavenumbers (k_-, k_+).*

We exclude the translation symmetries in time and space from the above multiplicity counting. Each defect depends smoothly on parameters in the nonlinearity and the diffusion matrix (see below).

Here, we say that a defect depends smoothly on wavenumbers and additional parameters μ if there exist smooth functions c_d and ω_d of (k_-, k_+, μ) and a family of defects $u_\mathrm{d}(x - c_\mathrm{d}t, \omega_\mathrm{d}t; k_-, k_+, \mu)$ such that u_d depends smoothly on (k_-, k_+, μ) as a function into $BC^1(\mathbb{R}\times\mathbb{R},\mathbb{R}^n)$ after transforming the argument $\xi = x - c_\mathrm{d}t$ according to

$$\xi \longmapsto \xi + \theta(\xi; k_-, k_+, \mu) \tag{1.12}$$

for an appropriate function θ for which $\theta'(\xi; k_-, k_+, \mu) \in BC^0(\mathbb{R})$ is smooth in (k_-, k_+, μ). The coordinate change in ξ is necessary in order to obtain continuity of the family since the wave trains depend continuously on k in BC^0 only after a rescaling $\xi \mapsto \xi/k$ that normalizes the spatial period.

1.3. Examples. To illustrate the theorem, we review the complex Ginzburg–Landau equation, whose defect solutions have been studied extensively in the literature, and the Burgers equation, which describes the dynamics of modulated wave trains.

Example I: The complex Ginzburg–Landau equation. The complex cubic Ginzburg–Landau equation (CGL) is given by

$$A_t = (1+\mathrm{i}\alpha)A_{xx} + A - (1+\mathrm{i}\beta)A|A|^2, \tag{1.13}$$

where the coefficients $\alpha, \beta \in \mathbb{R}$ are real, and where $x \in \mathbb{R}$, $t \geq 0$, and $A(x,t) \in \mathbb{C}$. The CGL has a family of wave trains given by

$$A(x,t) = A_{\mathrm{wt}}(kx - \omega t; k) = \sqrt{1-k^2}\,\mathrm{e}^{\mathrm{i}(kx-\omega t)}, \tag{1.14}$$

where the spatial wavenumber k and the temporal frequency ω are related via

$$\omega = \omega_{\mathrm{nl}}(k) = \beta + (\alpha - \beta)k^2.$$

Note that these waves exist only for $|k| < 1$. Due to the gauge invariance $A \mapsto \mathrm{e}^{\mathrm{i}\phi}A$ of the CGL, defects can actually be constructed, in a frame moving with the defect speed c_{d}, as heteroclinic and homoclinic orbits of an ODE. Indeed, coherent defects of the CGL are of the form

$$A(x,t) = A_{\mathrm{d}}(x - c_{\mathrm{d}}t, \omega_{\mathrm{d}}t) = a(x - c_{\mathrm{d}}t)\mathrm{e}^{\mathrm{i}\phi(x-c_{\mathrm{d}}t)}\mathrm{e}^{-\mathrm{i}\omega_{\mathrm{d}}t} \tag{1.15}$$

so that

$$A_{\mathrm{d}}(\xi,\tau) = a(\xi)\mathrm{e}^{\mathrm{i}[\phi(\xi)-\tau]}.$$

Substituting this ansatz into (1.13), it follows [22, Appendix B] that (a,z) satisfies the ODE

$$\begin{aligned} a_\xi &= a\,\mathrm{Re}\,z, \\ z_\xi &= -z^2 - \frac{1}{1+\mathrm{i}\alpha}\left[1 + \mathrm{i}\omega_{\mathrm{d}} - (1+\mathrm{i}\beta)a^2 + c_{\mathrm{d}}z\right], \end{aligned} \tag{1.16}$$

where $z = a_\xi/a + \mathrm{i}\phi_\xi$. Thus, defects of the CGL correspond to heteroclinic orbits of the ODE (1.16), while the equilibria connected by these orbits correspond to wave trains of the CGL. The observation made in [41] is that the dimensions of the unstable and stable manifolds of these equilibria are related to the group velocities of the corresponding wave trains: In a frame moving with the speed of the defect, the dimension of stable manifolds associated with wave trains that transport to the left is larger than that of wave trains that transport to the right. Equivalently, the dimension of unstable manifolds associated with wave trains that transport to the right is larger than that of wave trains that transport to the left. These dimensions, however, are directly related, via counting arguments, to the codimension of connecting orbits between equilibria as these arise as intersections of stable and unstable manifolds. This argument therefore establishes a beautiful connection between the intuition given by the group velocity and rigorous counts of the codimension of defects.

Equation (1.16) has been studied thoroughly in the literature (see, for instance, [41, 2, 10, 21, 22] for references). In particular, the CGL has been shown to admit sources (the so-called Nozaki–Bekki holes), sinks, and transmission defects (commonly referred to as homoclons [21]). We refer to [10] for existence results of these defects and to [28] for the stability of sinks. We will show in section 7 that the cubic-quintic Ginzburg–Landau equation (CQGL) admits contact defects.

Example II: The viscous Burgers equation. Arguably, the simplest possible defects are those with "small amplitude" that serve as interfaces between two wave trains with almost the same wavenumber. The term "small amplitude" therefore refers to the small difference of the asymptotic wavenumbers and not to the actual amplitudes of wave trains and defects which could be large.

One way of finding such defects is to derive an equation that describes slowly varying wavenumber modulations (see Figure 1.3) of the wave trains $u_{\mathrm{wt}}(kx - \omega_{\mathrm{nl}}(k)t; k)$. Hence, we fix a wavenumber k and seek solutions to the reaction-diffusion system (1.1) of the form

$$u(x,t) = u_{\mathrm{wt}}(kx - \omega_{\mathrm{nl}}(k)t + \Phi(X,T); k + \varepsilon\partial_X\Phi(X,T)) + \mathrm{O}(\varepsilon^2), \tag{1.17}$$

where $0 < \varepsilon \ll 1$, and where the variables (X,T) depend on (x,t) via

$$X = \varepsilon(x - c_{\mathrm{g}}t), \qquad T = \frac{1}{2}\varepsilon^2 t. \tag{1.18}$$

Thus, the function $q(X,T) := \partial_X\Phi(X,T)$ describes the slowly varying modulation of the wavenumber. It can be shown [26, 11] that $q(X,T)$ satisfies the viscous Burgers equation

$$\partial_T q = \lambda''_{\mathrm{lin}}(0)\,\partial_X^2 q - \omega''_{\mathrm{nl}}(k)\,\partial_X(q^2) \tag{1.19}$$

over time scales of order $\mathrm{O}(1)$ in T and that any solution to (1.19) yields, in fact, a solution to the reaction-diffusion system (1.1) via (1.17). Using (1.18), this validity result justifies the interpretation of the group velocity as the speed of propagation of small localized wave-package perturbations.

To analyze defects, suppose that the dispersion relation is convex near the wavenumber k so that $\omega''(k) > 0$ (the only difference for concave dispersion relations is that some of the signs below may change). Equation (1.19) admits stationary fronts of the form

$$q_{\mathrm{d}}(X) = -\sqrt{\frac{\check{\omega}}{\omega''_{\mathrm{nl}}(k)}}\tanh\left(\frac{\sqrt{\check{\omega}\omega''_{\mathrm{nl}}(k)}}{\lambda''_{\mathrm{lin}}(0)}X\right) \tag{1.20}$$

that converge to the equilibria $q_\pm = \mp\sqrt{2\check{\omega}/\omega''_{\mathrm{nl}}(k)}$ as $X \to \pm\infty$ for each $\check{\omega} > 0$. These fronts of (1.19) correspond to defects of (1.1) that connect the wave train with wavenumber $k + \varepsilon q_-$ to the wave train with wavenumber $k + \varepsilon q_+$. Since the dispersion relation is locally convex, and since $q_- > 0$ and $q_+ < 0$, we see that the group velocity of $k + \varepsilon q_-$ is larger than c_{g}, while the group velocity of $k + \varepsilon q_+$ is smaller than c_{g}. Thus, the defects described by (1.20) are sinks whose characteristics on each side point toward the interface. This is, of course, in agreement with the fact that (1.20) describes the viscous Lax shocks of the Burgers equation.

Another consequence of the above discussion is that sources do not exist in the small-amplitude limit. Indeed, for $\omega''(k) > 0$, the only waves of the viscous Burgers equation that connect q_- to q_+ for $q_- < q_+$ are rarefaction waves.[4]

We remark that it has been proved in [11] that the shocks given in (1.20) persist as transverse defects of the reaction-diffusion equation (1.1); note that this requires a proof as the Burgers equation is valid only over finite time intervals.

Example III: Absolute and essential instability of pulses and fronts. Bifurcations from travelling waves provide another mechanism by which defects can be created. Imagine, for instance, a pulse travelling with a nonzero speed through a stable spatially homogeneous background. Now, envision a scenario where the stable background becomes unstable through a Turing or Hopf instability upon varying appropriate external parameters: At a supercritical Turing bifurcation, stationary wave trains with small amplitude and nonzero wavenumber will bifurcate from the homogeneous background state. We proved in [44] that, under certain generic assumptions, modulated pulses arise that travel through these wave trains with nonzero speed. Since the wave trains will have zero group velocity, the modulated pulse is an elementary transverse transmission defect. Linear and nonlinear stability of these defects have been addressed in [45] and [17], respectively. Analogous bifurcations can also occur at fronts when the rest state ahead of the front destabilizes [48]. Last, we proved in [51] that both flip-flops and one-dimensional target patterns (see Figure 1.2(i) and (ii)) will bifurcate from standing pulses whose background states undergo a Hopf instability. This appears to be the mechanism that creates the chemical flip-flops observed in [38].

Plan of the paper. The paper is organized as follows. In section 2, we review robustness and stability properties of localized pulses and fronts that connect spatially homogeneous equilibria. The purpose of this review is to prepare the spatial-dynamics viewpoint that we shall adopt when we investigate defects. Section 3 is devoted to wave trains and their stability for both the temporal and the spatial-dynamical system. In section 4, we prepare the ground for the proof of Theorem 1.4 by investigating the spectra of the period maps associated with the linearization about defects. We then prove Theorem 1.4 in section 5 by linking robustness properties of defects to geometric transversality conditions of the spatial-dynamical system. We also show that the resulting geometric conditions are equivalent to the minimal-spectrum assumption. In section 6, we address stability, interactions, and bifurcations of defects as well as the influence of boundaries and inhomogeneities on their dynamics. Last, in section 7, we prove that the CQGL has contact defects in appropriate parameter regimes.

2. Localized travelling waves.

To illustrate the main ideas behind Theorem 1.4 and its proof, we review travelling waves that approach stable spatially homogeneous rest states. We describe both rigidly propagating travelling waves $u = u_{\mathrm{tw}}(x - c_{\mathrm{tw}}t)$ and oscillatory modulated

[4] Rarefaction waves appear as stationary fronts of (1.19) if the self-similarity scaling symmetry is exploited, which is respected by (1.19) but, in general, not by (1.1).

waves $u = u_{\mathrm{mtw}}(x - c_{\mathrm{mtw}}t, \omega_{\mathrm{mtw}}t)$ of reaction-diffusion systems

$$u_t = Du_{xx} + f(u), \qquad x \in \mathbb{R}, \tag{2.1}$$

with $u \in \mathbb{R}^n$.

2.1. Pulses and fronts.

Suppose that $u_{\mathrm{tw}}(x - ct)$ is a front that satisfies (2.1) and that connects two stable spatially homogeneous equilibria $u_\pm$ of (2.1) so that $u_{\mathrm{tw}}(\xi) \to u_\pm$ as $\xi \to \pm\infty$. Thus, we use the independent variables $(\xi, t) = (x - ct, t)$ so that (2.1) becomes

$$u_t = Du_{\xi\xi} + cu_\xi + f(u), \qquad \xi \in \mathbb{R}, \tag{2.2}$$

and $u_{\mathrm{tw}}(\xi)$ is an equilibrium. The linearization of (2.2) about the front $u_{\mathrm{tw}}(\xi)$ is given by

$$\mathcal{L}_{\mathrm{tw}}u = Du_{\xi\xi} + cu_\xi + f'(u_{\mathrm{tw}}(\xi))u, \tag{2.3}$$

which defines a closed unbounded operator $\mathcal{L}_{\mathrm{tw}}$ on $L^2(\mathbb{R}, \mathbb{R}^n)$ and on $BC^0(\mathbb{R}, \mathbb{R}^n)$. We shall see below that $u'_{\mathrm{tw}}(\xi)$ decays exponentially to zero as $|\xi| \to \infty$. As a consequence, $\lambda = 0$ is always an eigenvalue of $\mathcal{L}_{\mathrm{tw}}$ with eigenfunction $u'_{\mathrm{tw}}(\xi)$. This eigenvalue occurs due to the translation symmetry of (2.2) which implies that $u_{\mathrm{tw}}(\cdot + \xi_0)$ is a solution for each fixed spatial shift $\xi_0 \in \mathbb{R}$. Since $\mathcal{L}_{\mathrm{tw}}$ is Fredholm with index zero [24], we can apply Lyapunov–Schmidt reduction to the steady-state equation associated with (2.2) near the family of fronts. Exploiting the translation symmetry $\xi \mapsto \xi + \xi_0$, it is not difficult to see that fronts and pulses are robust with respect to small perturbations of the nonlinearity provided $\lambda = 0$ has geometric and algebraic multiplicity one as an eigenvalue of $\mathcal{L}_{\mathrm{tw}}$ (see [48] and the references therein for details).

An alternative approach to this problem is as follows. We seek travelling-wave solutions $u_{\mathrm{tw}}(x - ct)$ of (2.1). Substituting this ansatz gives the travelling-wave equation

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= -D^{-1}[cv + f(u)] \end{aligned} \tag{2.4}$$

in the $2n$-dimensional phase space, which we also write as

$$\mathbf{u}' = \mathcal{F}(\mathbf{u}; c),$$

where $\mathbf{u} = (u, v) \in \mathbb{R}^{2n}$. The front $u_{\mathrm{tw}}(\xi)$ which connects two stable spatially homogeneous equilibria $u_\pm$ corresponds to a heteroclinic orbit $\mathbf{u}_{\mathrm{tw}}(\xi)$ of (2.4) which connects the equilibria $\mathbf{u}_\pm = (u_\pm, 0)$. The eigenvalue problem for the operator $\mathcal{L}_{\mathrm{tw}}$ can now be written as the linear ODE

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= -D^{-1}[cv + f'(u_{\mathrm{tw}}(\xi))u - \lambda u]. \end{aligned} \tag{2.5}$$

For λ close to zero, this equation is close to the ODE linearization of (2.4) about the travelling wave $\mathbf{u}_{\mathrm{tw}}$. We can therefore expect that the spectral properties of $\mathcal{L}_{\mathrm{tw}}$ for λ close to zero are related to properties of the heteroclinic orbit of the travelling-wave ODE (2.4).

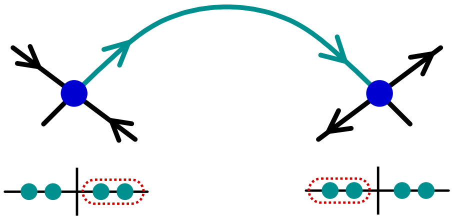


Figure 2.1. *A front with stable homogeneous background states corresponds to a heteroclinic orbit that connects saddles whose unstable and stable manifolds have dimension n. The insets show the eigenvalues of the linearization of* (2.4) *about the saddles.*

First, consider the PDE linearization about a homogeneous equilibrium $u_{\mathrm{tw}}(\xi) \equiv u_\pm$. Using Fourier transform, it is easy to see that λ belongs to the spectrum if and only if (2.5) has a nontrivial bounded solution. For $\lambda \gg 1$, (2.5) is close to $u_{\xi\xi} = \lambda D^{-1} u$, which is a linear hyperbolic equation with n stable and n unstable spatial eigenvalues $\nu = \pm\sqrt{\lambda/d_j}$, where $D = \mathrm{diag}(d_j)$. If we therefore assume that the homogeneous equilibria $u_\pm$ are spectrally stable, we can conclude that the corresponding two equilibria $\mathbf{u}_\pm = (u_\pm, 0)$ of the travelling-wave ODE (2.4) are saddles whose stable and unstable manifolds have equal dimension n. In particular, fronts and pulses with stable background states correspond to heteroclinic and homoclinic orbits to saddles with unstable and stable dimension equal to n (see Figure 2.1).

Next, robustness of fronts corresponds to robustness of the heteroclinic connection $\mathbf{u}_{\mathrm{tw}} = (u_{\mathrm{tw}}, u'_{\mathrm{tw}})$ as a solution to the ODE (2.4). In fact, heteroclinic orbits are robust as solutions in the phase space if and only if stable and unstable manifolds intersect transversely upon varying the parameter c near the wave speed c_{tw} of the front or pulse; if we add the equation $c_\xi = 0$ for the parameter c to (2.4) and denote the center-stable and center-unstable manifolds of the equilibria $\mathbf{u}_\pm$ by $W^{\mathrm{cu}}_{\mathrm{ext}}(\mathbf{u}_-)$ and $W^{\mathrm{cs}}_{\mathrm{ext}}(\mathbf{u}_+)$, then robustness is equivalent to requiring

$$T_{(\mathbf{u}_{\mathrm{tw}}, c_{\mathrm{tw}})} W^{\mathrm{cu}}_{\mathrm{ext}}(\mathbf{u}_-) + T_{(\mathbf{u}_{\mathrm{tw}}, c_{\mathrm{tw}})} W^{\mathrm{cs}}_{\mathrm{ext}}(\mathbf{u}_+) = \mathbb{R}^{2n} \times \mathbb{R}.$$

This, in turn, is equivalent to a minimal intersection in $\mathbb{R}^{2n}$ for fixed $c = c_{\mathrm{tw}}$,

$$T_{\mathbf{u}_{\mathrm{tw}}} W^{\mathrm{u}}(\mathbf{u}_-) \cap T_{\mathbf{u}_{\mathrm{tw}}} W^{\mathrm{s}}(\mathbf{u}_+) = \mathrm{span}\, \mathbf{u}'_{\mathrm{tw}}, \tag{2.6}$$

together with a Melnikov condition

$$M = \int_{-\infty}^{\infty} \langle \boldsymbol{\psi}(\xi), \partial_c \mathcal{F}(\mathbf{u}_{\mathrm{tw}}(\xi); c_{\mathrm{tw}}) \rangle \, \mathrm{d}\xi \neq 0. \tag{2.7}$$

Here, $\boldsymbol{\psi}$ denotes the unique (up to scalar multiples) nontrivial bounded solution of the adjoint variational equation

$$\begin{aligned} u_\xi &= f'(u_{\mathrm{tw}}(\xi))^* D^{-1} v, \\ v_\xi &= -u + cD^{-1} v. \end{aligned} \tag{2.8}$$

Note that (2.6) holds precisely when $\lambda = 0$ has geometric multiplicity one as an eigenvalue of $\mathcal{L}_{\mathrm{tw}}$, while (2.7) holds exactly when its algebraic multiplicity is one. The relation between the

ODE and the functional-analytic robustness argument becomes clearer when we approach the eigenvalue problem from an ODE viewpoint. Eigenfunctions correspond to bounded solutions of (2.3), while generalized eigenfunctions are found as bounded solutions of the derivative of the variational equation with respect to λ, evaluated in the eigenfunction.

To find bounded solutions of (2.3), we denote by $E^{\mathrm{s}}_+(\lambda)$ the λ-dependent linear subspace of initial conditions at $\xi = 0$ that lead to bounded solutions of (2.3) as $x \to \infty$ and by $E^{\mathrm{u}}_-(\lambda)$ the λ-dependent linear subspace that leads to bounded solutions as $x \to -\infty$. Using exponential dichotomies, we see that both subspaces are n-dimensional and are given as ranges of analytic families of projections $P^{\mathrm{s}}_+(\lambda)$ and $P^{\mathrm{u}}_-(\lambda)$. We may now choose analytic bases $\mathbf{e}^{\mathrm{u}}_j$ in $E^{\mathrm{u}}_-(\lambda)$ and $\mathbf{e}^{\mathrm{s}}_j$ in $E^{\mathrm{s}}_+(\lambda)$ and define the Evans function

$$\mathcal{E}(\lambda) := \det\left[\mathbf{e}^{\mathrm{s}}_1, \ldots, \mathbf{e}^{\mathrm{s}}_n, \mathbf{e}^{\mathrm{u}}_1, \ldots, \mathbf{e}^{\mathrm{u}}_n\right].$$

In particular, we have $\mathcal{E}(\lambda) = 0$ if and only if λ is an eigenvalue of $\mathcal{L}_{\mathrm{tw}}$. Furthermore, we have $\mathcal{E}(0) = 0$, and, upon expanding the determinant, it is also not hard to see that $\mathcal{E}'(0) \neq 0$ if and only if both (2.6) and (2.7) are met.

A related way to solve the eigenvalue problem in a neighborhood of $\lambda = 0$ consists of finding roots of the injection map

$$\iota(\lambda): \quad E^{\mathrm{u}}_-(\lambda) \times E^{\mathrm{s}}_+(\lambda) \longrightarrow \mathbb{C}^{2n}, \qquad (\mathbf{u}^-, \mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+. \tag{2.9}$$

Near $\lambda = 0$, this map can be pulled back to

$$\iota_0(\lambda): \quad E^{\mathrm{u}}_-(0) \times E^{\mathrm{s}}_+(0) \longrightarrow \mathbb{C}^{2n}, \qquad (\mathbf{u}^-, \mathbf{u}^+) \longmapsto P^{\mathrm{u}}_-(\lambda)\mathbf{u}^- - P^{\mathrm{s}}_+(\lambda)\mathbf{u}^+. \tag{2.10}$$

Note that $\iota_0(0)$ is Fredholm with index zero and null space $E^{\mathrm{u}}_-(0) \cap E^{\mathrm{s}}_+(0)$. If we denote the dimension of this null space by ℓ, which coincides with the geometric multiplicity of $\lambda = 0$, then we can compute the roots of ι_0 near $\lambda = 0$ via Lyapunov–Schmidt reduction, which results in a linear system of ℓ equations in ℓ variables. The λ-dependent determinant of the reduced equation is a reduced Evans function $\mathcal{E}_0(\lambda)$ whose roots, counted with multiplicity, coincide with those of $\mathcal{E}(\lambda)$ in a neighborhood of zero. In particular, these two functions differ only by a nonzero analytic factor. In our case, the geometric multiplicity of $\lambda = 0$ is one so that $\ell = 1$. Thus, the reduced Evans function $\mathcal{E}_0(\lambda)$ is a scalar function, and we have $\mathcal{E}_0'(0) \neq 0$ if and only if the algebraic multiplicity of $\lambda = 0$ is one.

This latter approach of the construction of a *reduced Evans function* for eigenvalue problems is the one that we shall adopt below.

2.2. Modulated waves.

Hopf bifurcations from rigidly propagating travelling waves lead to modulated waves $u_{\mathrm{mtw}}(x - ct, \omega t)$ for some temporal frequency $\omega = \omega_{\mathrm{mtw}}$ and an average speed of propagation $c = c_{\mathrm{mtw}}$, where the profile $u_{\mathrm{mtw}}(\xi, \tau)$ is 2π-periodic in its second argument. In particular, $u_{\mathrm{mtw}}(\xi, \tau)$ is a periodic orbit with period 2π of the reaction-diffusion system (2.1) in a comoving frame:

$$\omega u_\tau = D u_{\xi\xi} + c u_\xi + f(u). \tag{2.11}$$

As in the previous section, we assume that $u_{\mathrm{mtw}}(\xi, \tau)$ converges to two asymptotically stable spatially homogeneous equilibria $u_\pm$ of (2.1) as $\xi \to \pm\infty$, uniformly in τ. In other words, we require $u_{\rm mtw}(\xi,\cdot) \to u_\pm$ as $\xi \to \pm\infty$. We shall see later that this convergence is necessarily exponential.

To analyze robustness and stability of modulated waves, we consider the linearized period map $\Phi_{\rm mtw}$ that maps initial data $u(\cdot,0)$ to the solution $u(\cdot,2\pi)$ at time $\tau = 2\pi$ of

$$\omega u_\tau = D u_{\xi\xi} + c u_\xi + f'(u_{\rm mtw}(\xi,\tau))u. \tag{2.12}$$

The operator $\Phi_{\rm mtw} - 1$ is Fredholm with index zero when posed on $L^2(\mathbb{R},\mathbb{R}^n)$. Note that $\rho = 1$ is an eigenvalue of $\Phi_{\rm mtw}$ with geometric multiplicity equal to at least two since both $\partial_\xi u_{\rm mtw}$ and $\partial_\tau u_{\rm mtw}$ contribute one dimension each to the eigenspace. Using Lyapunov–Schmidt reduction and eliminating the space and time translational symmetries, we see that modulated waves are robust provided $\rho = 1$ has algebraic multiplicity two.

Alternatively, we may investigate modulated waves by casting (2.12) as the dynamical system

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega\partial_\tau u - cv - f(u)] \end{aligned} \tag{2.13}$$

in the ξ-variable. Modulated waves satisfy (2.13), which we can view as an abstract differential equation

$$\mathbf{u}' = \mathcal{F}(\mathbf{u}; c, \omega) \tag{2.14}$$

on the phase space $Y = H^{1/2}(S^1,\mathbb{R}^n) \times L^2(S^1,\mathbb{R}^n)$ of 2π-periodic functions. Note that the initial-value problem associated with (2.14) is ill-posed as can be readily seen by solving it using Fourier series for $f \equiv 0$ and $c = 0$. Despite this, we proved in [49] that the stable and unstable manifolds of the equilibria $(u_+,0)$ and $(u_-,0)$ can be constructed for (2.14) near the modulated-wave solution $\mathbf{u}_{\rm mtw} = (u_{\rm mtw}, \partial_\xi u_{\rm mtw})$. In this framework, robustness is again equivalent to the transverse crossing of the extended stable and unstable manifolds

$$T_{(\mathbf{u}_{\rm mtw},c_{\rm mtw},\omega_{\rm mtw})} W^{\rm cu}_{\rm ext}(\mathbf{u}_-) + T_{(\mathbf{u}_{\rm mtw},c_{\rm mtw},\omega_{\rm mtw})} W^{\rm cs}_{\rm ext}(\mathbf{u}_+) = Y \times \mathbb{R}^2$$

in the two-dimensional parameter (c,ω). As before, this is equivalent [49] to the statement that

$$\dim\left[T_{\mathbf{u}_{\rm mtw}} W^{\rm u}(\mathbf{u}_-) \cap T_{\mathbf{u}_{\rm mtw}} W^{\rm s}(\mathbf{u}_+)\right] = 2, \tag{2.15}$$

where the intersection is spanned by $\partial_\xi \mathbf{u}_{\rm mtw}$ and $\partial_\tau \mathbf{u}_{\rm mtw}$, together with the requirement that

$$\begin{aligned} &\det\,(M_{ij})_{i=1,2,\ j=c,\omega} \\ &\quad = \det\left(\int_{-\infty}^{\infty} \langle \boldsymbol{\psi}_i(\xi), \partial_j \mathcal{F}(\mathbf{u}_{\rm mtw}(\xi); c_{\rm mtw}, \omega_{\rm mtw})\rangle_Y \,\mathrm{d}\xi\right)_{i=1,2,\ j=c,\omega} \neq 0, \end{aligned} \tag{2.16}$$

where $\boldsymbol{\psi}_i$ are the two linearly independent bounded solutions of the adjoint variational equation

$$\begin{aligned} u_\xi &= \left[f'(u_{\rm mtw}(\xi,\tau))^* + \omega_{\rm mtw}\partial_\tau\right] D^{-1} v, \\ v_\xi &= -u + c_{\rm mtw} D^{-1} v. \end{aligned} \tag{2.17}$$

To elucidate the relation between the analytical and the geometric approaches outlined above, consider the eigenvalue problem $\Phi_{\mathrm{mtw}} u = \rho u$ for the linear period map Φ_{mtw} which is equivalent to the equation

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_{\mathrm{mtw}} \partial_\tau u - c_{\mathrm{mtw}} v - f'(u_{\mathrm{mtw}}(\xi,\tau))u + \lambda u] \end{aligned} \tag{2.18}$$

on Y. Here, we have used the Floquet ansatz $(u,v) \mapsto \mathrm{e}^{\lambda t}(u,v)$, where λ is the Floquet exponent and $\rho = \exp(2\pi\lambda/\omega_{\mathrm{mtw}})$ the associated Floquet multiplier which corresponds to an eigenvalue of Φ_{mtw}. We can now construct stable and unstable subspaces for (2.18) that depend analytically on λ for λ close to zero [49]. Both subspaces are infinite-dimensional, whence it is not obvious how to construct an Evans function using determinants as in the case of travelling waves.

One way is to employ Galerkin approximations and to replace $f'(u_{\mathrm{mtw}})$ by $P_m f'(u_{\mathrm{mtw}})$, where P_m is the orthogonal projection in L^2 onto the first m temporal Fourier modes. Following [49], it is not difficult to see that roots of the approximate Evans functions that are defined for the $2mn$-dimensional Fourier approximation converge with multiplicity as $m \to \infty$. In particular, for $m \gg 1$, winding-number calculations for the complex-analytic approximate Evans function give the correct number of eigenvalues for the full problem (2.18) on open bounded regions of the complex plane that do not intersect the absolute spectrum [47].

If we are only interested in computing roots locally, we can proceed as in the previous section and use the injection maps $\iota(\lambda)$ and $\iota_0(\lambda)$ that we defined in (2.9) and (2.10). The results in [49] imply in particular that the injection maps are again Fredholm with index zero. Using Lyapunov–Schmidt reduction, we see that eigenvalues can be computed locally as roots of a reduced determinant $\mathcal{E}_0(\lambda)$, which we may refer to as the reduced Evans function. For modulated waves, we then have $\mathcal{E}_0(0) = \mathcal{E}_0'(0) = 0$ and $\mathcal{E}_0''(0) \neq 0$ provided both (2.15) and (2.16) are satisfied.

3. Wave trains, group velocities, and spatial dynamics.

We are interested in defects that are spatially asymptotic to wave trains instead of to homogeneous steady states. In contrast to the exponentially stable homogeneous equilibria, wave trains always have a neutral mode, associated with their phase, and a resulting group velocity.

When interpreted in terms of spatial-dynamical systems, defects correspond to heteroclinic orbits of the modulated-wave equation (2.13) that connect two periodic orbits instead of two hyperbolic equilibria. The asymptotic periodic orbits have again a neutral direction. We will prove in this section that, after eliminating the neutral eigenvalue, the unstable dimension, and therefore the Fredholm index of the injection maps ι and ι_0 that we defined in the previous section, depends only on whether the group velocity of the wave train is larger or smaller than the speed of the defect. This result therefore allows us to relate group velocities of wave trains and Morse indices of the spatial-dynamical system.

We refer to [53, 45, 11] for more details and references concerning the material in this section.

3.1. Existence and stability of wave trains.

We assume that, for some nonzero temporal frequency ω_0 and a certain spatial wavenumber k_0, there exists a nonconstant wave-train solution $u(x,t) = u_{\mathrm{wt}}(k_0 x - \omega_0 t)$ of (1.1), where $u_{\mathrm{wt}}(\phi)$ is 2π-periodic in its argument. Throughout this section, we focus on the case $k_0 \neq 0$ and discuss the somewhat simpler case $k_0 = 0$ in section 3.3 below.

Substituting the ansatz for $u(x,t)$ into (1.1), we see that $u_{\mathrm{wt}}(\phi)$ must be a 2π-periodic solution of the ODE

$$k_0^2 D\partial_{\phi\phi}u + \omega_0\partial_\phi u + f(u) = 0. \tag{3.1}$$

Linearizing this equation about u_{wt}, we obtain the linear operator $\mathcal{L}_{\mathrm{wt}}$,

$$\mathcal{L}_{\mathrm{wt}} := k_0^2 D\partial_{\phi\phi} + \omega_0\partial_\phi + f'(u_{\mathrm{wt}}(\phi)), \tag{3.2}$$

which defines a closed operator on $L^2(0,2\pi)$ with domain $H^2_{\mathrm{per}}(0,2\pi)$.

Hypothesis 3.1. *The origin $\lambda = 0$ is algebraically simple as an eigenvalue of $\mathcal{L}_{\mathrm{wt}}$ on $L^2(0,2\pi)$ with eigenfunction $\partial_\phi u_{\mathrm{wt}}$.*

Spectral stability of the wave train u_{wt} on $\mathbb{R}$ is determined as follows. A complex number $\check{\lambda}$ is in the spectrum of $\mathcal{L}_{\mathrm{wt}}$ considered as a closed operator on $L^2(\mathbb{R},\mathbb{C}^n)$ with domain $H^2(\mathbb{R},\mathbb{C}^n)$ if and only if there are a $\nu \in \mathrm{i}\mathbb{R}$ and a 2π-periodic function $w(\phi)$ such that $\mathcal{L}_{\mathrm{wt}}u = \check{\lambda}u$ for $x \in \mathbb{R}$, where

$$u(\phi) = \mathrm{e}^{\nu\phi/k_0}w(\phi). \tag{3.3}$$

Note that the resulting equation for w is

$$\check{\lambda}w = D\left[k_0\partial_\phi + \nu\right]^2 w + c_{\mathrm{p}}\left[k_0\partial_\phi + \nu\right]w + f'(u_{\mathrm{wt}}(\phi))w, \tag{3.4}$$

where $c_{\mathrm{p}} = \omega_0/k_0$ is the phase speed of the wave trains. Hypothesis 3.1 implies that there is an analytic function $\lambda_{\mathrm{lin}}(\nu)$ with $\lambda_{\mathrm{lin}}(0) = 0$ such that $\check{\lambda}$ close to zero is in the spectrum if and only if $\check{\lambda} = \lambda_{\mathrm{lin}}(\nu)$ for some $\nu \in \mathrm{i}\mathbb{R}$ close to zero.

Hypothesis 3.2. *The linear dispersion relation is dissipative so that $d_\| := \lambda''_{\mathrm{lin}}(0) > 0$. Furthermore, the spectrum of $\mathcal{L}_{\mathrm{wt}}$ on $L^2(\mathbb{R},\mathbb{C}^n)$ lies in the open left half-plane except for the spectrum near $\check{\lambda} = 0$, which is captured by the linear dispersion relation $\check{\lambda} = \lambda_{\mathrm{lin}}(\nu)$ with $\nu \in \mathrm{i}\mathbb{R}$ close to zero.*

The coefficient $d_\|$ measures the effective diffusion rate of perturbations in the direction of propagation of the wave train (see (1.19)). Planar wave trains in $x \in \mathbb{R}^2$ have an additional diffusion coefficient $d_\perp$ that measures diffusive decay of perturbations transverse to the direction of propagation.

Hypothesis 3.1 implies that there exists a family $u_{\mathrm{wt}}(kx - \omega t; k)$ of wave trains, defined for k close to k_0, which are stable and whose frequency is given by a smooth nonlinear dispersion relation $\omega = \omega_{\mathrm{nl}}(k)$ with $\omega_{\mathrm{nl}}(k_0) = \omega_0$.

Hypothesis 3.3. *We assume that the nonlinear dispersion relation is genuinely nonlinear, which means that $\omega''_{\mathrm{nl}}(k_0) \neq 0$.*

We denote the phase and group velocities by

$$c_{\mathrm{p}} = \frac{\omega_{\mathrm{nl}}(k)}{k}, \qquad c_{\mathrm{g}} = \frac{\mathrm{d}\omega_{\mathrm{nl}}(k)}{\mathrm{d}k}. \tag{3.5}$$

Using these definitions, it turns out [53, 11] that the Taylor series of the linear dispersion relation $\lambda_{\mathrm{lin}}(\nu)$ at $\nu = 0$ is given by

$$\lambda_{\mathrm{lin}}(\nu) = \left[c_{\mathrm{p}} - c_{\mathrm{g}}\right]\nu + d_\|\nu^2 + \mathrm{O}(\nu^3). \tag{3.6}$$

3.2. Spectra of wave trains in different frames. We discuss the dependence of the linear dispersion relation on the frame in which it is computed. This issue will play a crucial role below. Consider the reaction-diffusion equation (1.1) in a frame moving with an arbitrary, but fixed, speed c_d. In the variable $\xi = x - c_\mathrm{d}t$, we get

$$\phi = k_0x - \omega_0 t = k_0\xi - (\omega_0 - k_0c_\mathrm{d})t.$$

Thus, we set $\omega_\mathrm{d} = \omega_0 - k_0c_\mathrm{d}$ and define $\tau = \omega_\mathrm{d}t$. In the (ξ, τ) coordinates, (1.1) becomes

$$\omega_\mathrm{d}u_\tau = Du_{\xi\xi} + c_\mathrm{d}u_\xi + f(u), \tag{3.7}$$

and the wave trains are time-periodic solutions $u(\xi, \tau) = u_\mathrm{wt}(k_0\xi - \tau)$ with period 2π in τ.

Following section 2.2, we linearize the period map of (3.7) about the wave train so that

$$\Phi_\mathrm{wt} :\ u(\xi, 0) \longmapsto u(\xi, 2\pi)$$

is the solution map of the linear equation

$$\omega_\mathrm{d}u_\tau = Du_{\xi\xi} + c_\mathrm{d}u_\xi + f'(u_\mathrm{wt}(k_0\xi - \tau))u.$$

Note that the operator Φ_wt is not Fredholm on $L^2(\mathbb{R}, \mathbb{C}^n)$. Its spectrum can be computed as follows [45]. A nonzero number $\rho \in \mathbb{C}$ is in the spectrum of Φ_wt if and only if the linearized eigenvalue problem

$$\lambda u = Du_{\xi\xi} + c_\mathrm{d}u_\xi - \omega_\mathrm{d}u_\tau + f'(u_\mathrm{wt}(k_0\xi - \tau))u \tag{3.8}$$

has a bounded nonzero solution $u(\xi, \tau)$ that is 2π-periodic in τ, where the Floquet multiplier ρ and the Floquet exponent λ are related via $\rho = \exp(2\pi\lambda/\omega_\mathrm{d})$. These solutions can be calculated using the Floquet ansatz

$$u(\xi, \tau) = \mathrm{e}^{\nu\xi}\, w(k_0\xi - \tau), \tag{3.9}$$

where $w(\phi)$ is 2π-periodic and $\nu \in \mathrm{i}\mathbb{R}$. Upon substituting this ansatz into (3.8), we see after some calculations that w needs to satisfy the equation

$$[\lambda + (c_\mathrm{p} - c_\mathrm{d})\nu]\, w = D\, [k_0\partial_\phi + \nu]^2\, w + c_\mathrm{p}\, [k_0\partial_\phi + \nu]\, w + f'(u_\mathrm{wt}(\phi))w, \tag{3.10}$$

where $c_\mathrm{p} = \omega_0/k_0$ is the phase speed of the wave trains.

Comparing (3.10) with (3.4), we see that $\check{\lambda}$ is in the spectrum of the wave trains, computed in the frame moving with the phase speed c_p, if and only if

$$\lambda = \check{\lambda} + (c_\mathrm{d} - c_\mathrm{p})\nu \tag{3.11}$$

is a Floquet exponent of Φ_wt, computed in a frame moving with speed c_d, where $\nu \in \mathrm{i}\mathbb{R}$ is the associated spatial Floquet exponent. We denote by Σ_wt the set of all Floquet exponents λ of Φ_wt.

In particular, it follows from (3.6) that λ close to zero is in the Floquet spectrum of Φ_wt on $L^2(\mathbb{R}, \mathbb{C}^n)$, computed in the frame moving with speed c_d, if and only if

$$\lambda = \lambda_\mathrm{lin}(\nu) + (c_\mathrm{d} - c_\mathrm{p})\nu = [c_\mathrm{d} - c_\mathrm{g}]\,\nu + d_\parallel\nu^2 + \mathrm{O}(\nu^3) \tag{3.12}$$

for some $\nu \in \mathrm{i}\mathbb{R}$ close to zero. Note that $c_\mathrm{g} - c_\mathrm{d}$ represents the relative group velocity, i.e., the group velocity of the wave trains measured in the frame that moves with speed c_d.

3.3. Spatially homogeneous oscillations. In this section, we account for the differences that occur when the wavenumber of the wave trains vanishes. In other words, we consider spatially homogeneous oscillations $u(x,t) = u_{\mathrm{wt}}(-\omega_0 t)$, where $\omega_0 \neq 0$ and $u'_{\mathrm{wt}}(\phi)$ is not the zero function.

The spectrum of the period map Φ_{wt} associated with the spatially homogeneous oscillations can be computed easily. Indeed, Fourier transform in space reduces the time-periodic linearized parabolic equation

$$\omega_{\mathrm{d}} u_\tau = D u_{\xi\xi} + c_{\mathrm{d}} u_\xi + f'(u_{\mathrm{wt}}(-\tau))u \tag{3.13}$$

to the collection

$$\omega_{\mathrm{d}} \hat{u}_\tau = \nu^2 D \hat{u} + c_{\mathrm{d}} \nu \hat{u} + f'(u_{\mathrm{wt}}(-\tau))\hat{u} \tag{3.14}$$

of ODEs for purely imaginary Fourier exponents $\nu = \mathrm{i}\gamma$ with $\gamma \in \mathbb{R}$. Note that $\omega_{\mathrm{d}} = \omega_0$ since $k_0 = 0$. We denote by $\rho = \exp(2\pi\lambda/\omega_{\mathrm{d}})$ the complex temporal Floquet multipliers of the parabolic equation (3.13).

First, we need to replace Hypothesis 3.1 by the assumption that $\rho = 0$ is an algebraically simple multiplier for $\nu = 0$ and $c_{\mathrm{d}} = 0$. This allows us to continue the Floquet multiplier ρ as a smooth function $\rho = \rho(\nu)$ for any ν close to zero. The resulting dispersion relation for the temporal Floquet exponents is denoted by $\lambda(\nu)$, where $\lambda(0) = 0$. When $c_{\mathrm{d}} = 0$, ν enters only at quadratic order so that

$$\lambda = d_{\|}\nu^2 + \mathrm{O}(\nu^4).$$

Applying Fenichel's singular perturbation theory [16] to (3.1), it is straightforward to see that the spatially homogeneous oscillations are accompanied by a family of wave trains for wavenumbers k close to zero,[5] where wavenumber and frequency are related via a smooth nonlinear dispersion relation $\omega_{\mathrm{nl}}(k)$.

Next, we replace Hypothesis 3.2 by the assumption that $d_{\|} > 0$ and that the curve $\lambda(\nu)$, with ν close to the origin, captures all temporal Floquet exponents of the collection of ODEs (3.14) in the closed right half-plane $\mathrm{Re}\,\lambda \geq 0$. Last, we assume that Hypothesis 3.3 is met also when $k_0 = 0$. Inspecting the boundary-value problem

$$\lambda \hat{u} = D\nu^2 \hat{u} + c_{\mathrm{d}}\nu\hat{u} + f'(u_{\mathrm{wt}}(-\tau))\hat{u}, \qquad \hat{u}(0) = \hat{u}(2\pi), \tag{3.15}$$

for the temporal Floquet exponents λ, we see immediately that the dispersion relation for $c_{\mathrm{d}} \neq 0$ is related to the dispersion relation for $c_{\mathrm{d}} = 0$ via

$$\lambda = \lambda_{\mathrm{lin}}(\nu) + c_{\mathrm{d}}\nu = c_{\mathrm{d}}\nu + d_{\|}\nu^2 + \mathrm{O}(\nu^3),$$

which is the equivalent to (3.12) after formally setting $c_{\mathrm{g}} = 0$.

[5] In passing, we remark that we are not aware of a short direct proof of this fact that does not use Fenichel's theorem. Ginzburg–Landau approximations of the dynamics near homogeneous oscillations [56] capture waves with long wavelength but are, unfortunately, only valid over finite time intervals.

3.4. Spatial dynamics and relative Morse indices. We explore the implications of the results reviewed in the previous sections for the spatial-dynamical system associated with (3.7). Thus, we write (3.7) as

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_{\mathrm{d}}\partial_\tau u - c_{\mathrm{d}} v - f(u)], \end{aligned} \tag{3.16}$$

where $\mathbf{u} = (u, v) \in Y = H^{1/2}(S^1, \mathbb{R}^n) \times L^2(S^1, \mathbb{R}^n)$. One of the key features of (3.16) that we shall exploit over and over again is its equivariance with respect to the S^1-symmetry

$$\Gamma_\theta : \quad Y \longrightarrow Y, \quad (u, v)(\tau) \longmapsto (u, v)(\tau - \theta), \qquad \theta \in S^1, \tag{3.17}$$

that is induced by the temporal time shift. Note that the wave trains $u_{\mathrm{wt}}(k\xi - \tau; k)$ of (1.1) correspond to periodic orbits

$$\mathbf{u}_{\mathrm{wt}}(\xi) = (u_{\mathrm{wt}}(k\xi - \cdot; k), k\partial_\phi u_{\mathrm{wt}}(k\xi - \cdot; k))$$

of (3.16) with period $2\pi/k$ which are, in fact, relative equilibria with respect to the temporal time-shift symmetry. The eigenvalue problem (3.8) becomes

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_{\mathrm{d}}\partial_\tau u - c_{\mathrm{d}} v - f'(u_{\mathrm{wt}}(k\xi - \tau; k))u + \lambda u]. \end{aligned} \tag{3.18}$$

For each fixed value of λ, we say that ν is a spatial Floquet exponent of (3.18) if there is a 2π-periodic function $\mathbf{w}(\phi)$ with values in Y so that $\mathbf{u}(\xi) = \mathrm{e}^{\nu\xi}\mathbf{w}(k\xi)$ is a solution of (3.18). In fact, $\mathbf{w}(k\xi)$ will be of the form $[\mathbf{w}(k\xi)](\tau) = \check{\mathbf{w}}(k\xi - \tau)$, where the first component of $\check{\mathbf{w}}$ satisfies (3.10). Note that the spatial Floquet exponents ν are not unique, so we should restrict their imaginary parts to $0 \le \operatorname{Im}\nu < k$.

Spatial Floquet theory [34, 49] implies the following facts. For each fixed λ, there are projections $P^j_{\mathrm{wt}}(\xi; \lambda) \in \mathrm{L}(Y)$, labeled by $j = \mathrm{c, s, u}$, with the following properties. The projections are $2\pi/k$-periodic and strongly continuous in ξ, and their sum is the identity. The ranges of the stable and unstable projections $P^{\mathrm{s}}_{\mathrm{wt}}(\xi_0; \lambda)$ and $P^{\mathrm{u}}_{\mathrm{wt}}(\xi_0; \lambda)$ are infinite-dimensional and consist of all initial data at $\xi = \xi_0$ of solutions to (3.18) that decay exponentially for $\xi \to \infty$ and $\xi \to -\infty$, respectively. The range of the center projection $P^{\mathrm{c}}_{\mathrm{wt}}(\xi_0; \lambda)$ is finite-dimensional, and, for each initial value in $\mathrm{Rg}(P^{\mathrm{c}}_{\mathrm{wt}}(\xi_0; \lambda))$, the corresponding solution to (3.18) exists for $\xi \in \mathbb{R}$ and grows at most algebraically in ξ as $\xi \to \pm\infty$. The ranges of the projections $P^j_{\mathrm{wt}}(\xi; \lambda)$ can be obtained by taking the closure of the eigenfunctions $\mathbf{w}(k\xi) \in Y$, together with the associated generalized eigenfunctions if these exist, of all spatial Floquet exponents ν with $\operatorname{Re}\nu = 0$ for $j = \mathrm{c}$, $\operatorname{Re}\nu > 0$ for $j = \mathrm{u}$, and $\operatorname{Re}\nu < 0$ for $j = \mathrm{s}$.

The center projection $P^{\mathrm{c}}_{\mathrm{wt}}(\xi; \lambda)$ is nonzero if and only if λ is a temporal Floquet exponent of Φ_{wt}. In particular, $P^{\mathrm{c}}_{\mathrm{wt}}(\xi; \lambda) = 0$ for all λ with $\operatorname{Re}\lambda > 0$ since $|\rho| > 1$ belongs to the resolvent set of Φ_{wt}. In this case, the remaining projections $P^{\mathrm{s}}_{\mathrm{wt}}(\xi; \lambda)$ and $P^{\mathrm{u}}_{\mathrm{wt}}(\xi; \lambda)$ are analytic in λ. We are interested in counting the dimension of $\mathrm{Rg}(P^{\mathrm{u}}_{\mathrm{wt}}(\xi; \lambda))$ by comparing it to $\mathrm{Rg}(P^{\mathrm{u}}_{\mathrm{wt}}(0; \lambda_*))$, where λ_* is fixed so that $\operatorname{Re}\lambda_* \gg 1$ is positive.

Definition 3.4. *The* relative Morse index $i_{\mathrm{wt}}(\lambda)$ *is defined to be the Fredholm index of*

$$P^{\mathrm{u}}_{\mathrm{wt}}(\xi; \lambda) : \quad \mathrm{Rg}(P^{\mathrm{u}}_{\mathrm{wt}}(0; \lambda_*)) \longrightarrow \mathrm{Rg}(P^{\mathrm{u}}_{\mathrm{wt}}(\xi; \lambda)). \tag{3.19}$$

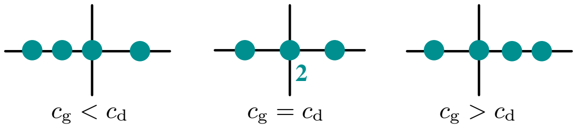


Figure 3.1. *The spatial Floquet spectrum of wave trains at $\lambda = 0$.*

We proved in [49] that the relative Morse index is well defined for all λ that are not temporal Floquet exponents of Φ_{wt}. Furthermore, it does not depend on ξ and on the choice of λ_* (as long as $\operatorname{Re}\lambda_* > 0$).

Hence, the relative Morse index $i_{\mathrm{wt}}(\lambda)$ is constant on each connected component of $\mathbb{C}\setminus\Sigma_{\mathrm{wt}}$, which corresponds to the resolvent set of Φ_{wt}. Note that we have $i_{\mathrm{wt}}(\lambda) = 0$ for all λ in the connected component of $\mathbb{C} \setminus \Sigma_{\mathrm{wt}}$ that contains the open right half-plane in $\mathbb{C}$. To compute Morse indices, we will use the following straightforward bordering lemma whose proof we shall omit.

Lemma 3.5. *Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces and that $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is a Fredholm operator with index $i(\mathcal{A})$. The operator*

$$\mathcal{S} = \begin{pmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{pmatrix} : \quad \mathcal{X} \times \mathbb{R}^p \longrightarrow \mathcal{Y} \times \mathbb{R}^q$$

is then Fredholm with index $i(\mathcal{S}) = i(\mathcal{A}) + p - q$ provided $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ are bounded and linear.

The following lemma predicts how the relative Morse index $i_{\mathrm{wt}}(\lambda)$ changes if we vary λ near $\lambda = 0$ (see Figure 3.1 for an illustration).

Lemma 3.6. *Assume that the hypotheses stated in section 3.1 are met. The spatial-dynamical system (3.18) has a geometrically simple spatial Floquet exponent $\nu = 0$ for $\lambda = 0$. The Floquet exponent $\nu = 0$ is simple if $c_{\mathrm{d}} \neq c_{\mathrm{g}}$, while it has algebraic multiplicity two if $c_{\mathrm{d}} = c_{\mathrm{g}}$. If $c_{\mathrm{d}} \neq c_{\mathrm{g}}$, then (3.18) has a simple spatial Floquet exponent $\nu = \nu(\lambda)$ for all λ close to zero and*

$$\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\bigg|_{\lambda=0} = \frac{1}{c_{\mathrm{d}} - c_{\mathrm{g}}}. \tag{3.20}$$

For $\lambda < 0$ close to zero, the relative Morse index $i_{\mathrm{wt}}(\lambda)$ is therefore $+1$ if $c_{\mathrm{g}} > c_{\mathrm{d}}$ and -1 if $c_{\mathrm{g}} < c_{\mathrm{d}}$.

Proof. The statement follows immediately from (3.12), the Cauchy–Riemann equations, and the bordering Lemma 3.5. ■

4. Spectral properties of defects.

We now turn to elementary defects $u_{\mathrm{d}}(\xi, \tau)$ that satisfy

$$\omega_{\mathrm{d}} u_\tau = D u_{\xi\xi} + c_{\mathrm{d}} u_\xi + f(u). \tag{4.1}$$

We are interested in the spectrum of the linear period map

$$\Phi_{\mathrm{d}} : \; u(\xi, 0) \longmapsto u(\xi, 2\pi)$$

associated with the linearization

$$\omega_{\mathrm{d}} u_\tau = D u_{\xi\xi} + c_{\mathrm{d}} u_\xi + f'(u_{\mathrm{d}}(\xi, \tau)) u$$

of (4.1) about the defect $u_\mathrm{d}(\xi,\tau)$. We denote by λ the Floquet exponents of Φ_d and by $\rho = \exp(2\pi\lambda/\omega_\mathrm{d})$ its Floquet multipliers, i.e., elements in its spectrum.

We adopt a dynamical-systems point-of-view and write (4.1) as the modulated-wave equation

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f(u)]. \end{aligned} \tag{4.2}$$

We write $\mathbf{u} = (u, v)$ and consider (4.2) on the space $Y = H^{1/2}(S^1,\mathbb{R}^n) \times L^2(S^1,\mathbb{R}^n)$. We shall also use the space $Y^1 = H^1(S^1,\mathbb{R}^n) \times H^{1/2}(S^1,\mathbb{R}^n)$. We say that $\mathbf{u}(\xi)$ is a *solution* of (4.2) on an interval $J \subset \mathbb{R}$ if $\mathbf{u}$ is contained in $L^2(J,Y^1) \cap H^1(J,Y)$ and it satisfies (4.2) in Y for all $\xi \in J$. Definition 1.1 implies that

$$\mathbf{u}_\mathrm{d}(\xi) = (u_\mathrm{d}(\xi,\cdot), \partial_\xi u_\mathrm{d}(\xi,\cdot))$$

satisfies (4.2) and that $\mathbf{u}_\mathrm{d}(\xi,\cdot) - \mathbf{u}_\mathrm{wt}(k_\pm\xi + \theta_\pm(\xi) - \cdot; k_\pm)$ converges to zero in Y^1 as $\xi \to \pm\infty$.

Throughout this section, we denote by $\Sigma^\pm_\mathrm{wt}$ the set of all temporal Floquet exponents λ of the asymptotic wave trains with wavenumber $k_\pm$, computed in the frame moving with speed c_d.

4.1. Exponential dichotomies.

We begin by analyzing the linear system

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f'(u_\mathrm{d}(\xi,\cdot))u + \lambda u], \end{aligned} \tag{4.3}$$

where the parameter λ represents potential temporal Floquet exponents of Φ_d.

Given one of the sets $J = \mathbb{R}^+$, $J = \mathbb{R}^-$, or $J = \mathbb{R}$, we say that (4.3) has an exponential dichotomy on J if there exist strongly continuous families $\{\Phi^\mathrm{s}(\xi,\zeta)\}_{\xi\geq\zeta,\,\xi,\zeta\in J}$ and $\{\Phi^\mathrm{u}(\xi,\zeta)\}_{\xi\leq\zeta,\,\xi,\zeta\in J}$ of operators in $\mathrm{L}(Y)$ as well as positive constants C and κ such that

(i) $\Phi^j(\xi,\sigma)\Phi^j(\sigma,\zeta) = \Phi^j(\xi,\zeta)$ for $j = \mathrm{s},\mathrm{u}$ and $\Phi^\mathrm{s}(\xi,\xi) + \Phi^\mathrm{u}(\xi,\xi) = 1$,

(ii) $\|\Phi^\mathrm{s}(\xi,\zeta)\| + \|\Phi^\mathrm{u}(\zeta,\xi)\| \leq C\mathrm{e}^{-\kappa|\xi-\zeta|}$,

(iii) $\Phi^\mathrm{s}(\xi,\zeta)\mathbf{u}_0$ and $\Phi^\mathrm{u}(\xi,\zeta)\mathbf{u}_0$ satisfy (4.3) for $\xi > \zeta$ and $\xi < \zeta$, respectively, for each $\mathbf{u}_0 \in Y$ provided $\xi,\zeta \in J$.

Thus, if a dichotomy[6] exists, then the operators $P^j(\xi) := \Phi^j(\xi,\xi)$ with $j = \mathrm{s},\mathrm{u}$ are complementary projections in Y. We denote their ranges at $\xi = 0$ by $E^\mathrm{s}(\lambda)$ and $E^\mathrm{u}(\lambda)$, where $E^\mathrm{s}(\lambda) \oplus E^\mathrm{u}(\lambda) = Y$.

Corollary A.2 in Appendix A.1 states that (4.3) has an exponential dichotomy on $J = \mathbb{R}^\pm$ if and only if the asymptotic equation

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f'(u_\mathrm{wt}(k_\pm\xi - \cdot; k_\pm))u + \lambda u] \end{aligned} \tag{4.4}$$

has an exponential dichotomy on $\mathbb{R}$. Furthermore, the projections $P^j_{\mathrm{wt},\pm}(\xi)$ of the asymptotic equation (4.4) and those of the linearization (4.3) about the defect differ only by a compact operator.

[6] The dichotomies and projections will depend on λ. We will suppress this dependence in our notation.

Thus, it remains to find out when the asymptotic equation has an exponential dichotomy on $\mathbb{R}$. The criterion in [34] for the existence of dichotomies to (4.4) is simply that it does not have purely imaginary Floquet exponents $\nu \in \mathrm{i}\mathbb{R}$, i.e., solutions of the form $\mathbf{u}(\xi) = \exp(\nu\xi)\mathbf{w}(k\xi)$ for some 2π-periodic function $\mathbf{w}$. Using the results reviewed in section 3, we can therefore conclude that (4.4) has an exponential dichotomy on $\mathbb{R}$ precisely when $\lambda \notin \Sigma^{\pm}_{\mathrm{wt}}$.

Last, we discuss what happens if the asymptotic equation does not have an exponential dichotomy. Since the Floquet multipliers of the asymptotic equation form a discrete set and accumulate at the origin [34], there are at most a finite number of them on the unit circle for any given λ. The idea is then to seek solutions to (4.3) of the form

$$\mathrm{u}(\xi) = \mathrm{e}^{\eta\xi}\check{\mathrm{u}}(\xi) \tag{4.5}$$

for a small nonzero weight η so that $\check{\mathrm{u}}(\xi)$ satisfies the equation

$$\begin{aligned} \check{u}_\xi &= -\eta\check{u} + \check{v}, \\ \check{v}_\xi &= -\eta\check{v} + D^{-1}[\omega_{\mathrm{d}}\partial_\tau\check{u} - c_{\mathrm{d}}\check{v} - f'(u_{\mathrm{d}}(\xi,\tau))\check{u} + \lambda\check{u}]. \end{aligned} \tag{4.6}$$

Suppose now that at least one spatial Floquet exponent $\nu = \mathrm{i}\gamma$ of (4.3) lies on the imaginary axis. A small exponential weight $\eta \neq 0$ will move this Floquet multiplier off the imaginary axis. Exploiting the conjugacy (4.5) between solutions to (4.6) and (4.3), we can construct center-stable dichotomies $\Phi^{\mathrm{cs}}(\xi,\zeta)$ and complementary strong-unstable dichotomies $\Phi^{\mathrm{uu}}(\xi,\zeta)$ for (4.3) by using the stable and unstable dichotomies of (4.6) for sufficiently small but nonzero weights $\eta > 0$. Analogously, upon using $\eta < 0$ close to zero, we find center-unstable and strong-stable dichotomies, $\Phi^{\mathrm{cu}}(\xi,\zeta)$ and $\Phi^{\mathrm{ss}}(\xi,\zeta)$, respectively, for (4.3). We can use these dichotomies to define a center projection

$$\begin{aligned} \mathrm{Rg}(\Phi^{\mathrm{c}}(\xi,\xi)) &:= \mathrm{Rg}(\Phi^{\mathrm{cs}}(\xi,\xi)) \cap \mathrm{Rg}(\Phi^{\mathrm{cu}}(\xi,\xi)), \\ \mathrm{N}(\Phi^{\mathrm{c}}(\xi,\xi)) &:= \mathrm{N}(\Phi^{\mathrm{cs}}(\xi,\xi)) + \mathrm{N}(\Phi^{\mathrm{cu}}(\xi,\xi)) \end{aligned}$$

and a corresponding center evolution $\Phi^{\mathrm{c}}(\xi,\zeta)$ for all $\xi,\eta \in J$. Since the unstable subspaces for $J = \mathbb{R}^+$ are arbitrary at $\xi = 0$ [39], we may also arrange that

$$\Phi^{\mathrm{cu}}(\xi,\zeta) = \Phi^{\mathrm{c}}(\xi,\zeta) + \Phi^{\mathrm{uu}}(\xi,\zeta), \qquad \Phi^{\mathrm{cs}}(\xi,\zeta) = \Phi^{\mathrm{c}}(\xi,\zeta) + \Phi^{\mathrm{ss}}(\xi,\zeta).$$

We use subscripts ± to distinguish the exponential dichotomies on $\mathbb{R}^+$ and on $\mathbb{R}^-$.

4.2. Fredholm indices. Suppose that λ is not a temporal Floquet exponent of either one of the asymptotic wave trains, i.e., $\lambda \notin \Sigma^-_{\mathrm{wt}} \cup \Sigma^+_{\mathrm{wt}}$. The discussion in the preceding section shows that (4.3) has exponential dichotomies on both $\mathbb{R}^+$ and $\mathbb{R}^-$. We denote by $E^{\mathrm{s}}_+(\lambda)$ the stable subspace at $\xi = 0$ of the dichotomy on $\mathbb{R}^+$ and by $E^{\mathrm{u}}_-(\lambda)$ the unstable subspace at $\xi = 0$ of the dichotomy on $\mathbb{R}^-$. Thus, we conclude that there exists a bounded solution to the linear equation (4.3) if and only if $E^{\mathrm{u}}_-(\lambda)$ and $E^{\mathrm{s}}_+(\lambda)$ intersect nontrivially.

For each $\lambda \notin \Sigma^-_{\mathrm{wt}} \cup \Sigma^+_{\mathrm{wt}}$, we therefore define the injection map

$$\iota(\lambda): \quad E^{\mathrm{u}}_-(\lambda) \times E^{\mathrm{s}}_+(\lambda) \longrightarrow Y, \qquad (\mathbf{u}^-,\mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+.$$

Since the evolutions $\Phi_\pm^{\mathrm{j}}$ with $j = \mathrm{s}, \mathrm{u}$ can be chosen to depend analytically on λ [49], the injection map ι is analytic in λ. Whenever $\iota(\lambda_*)$ is Fredholm of index zero with a nontrivial null space, we can use Lyapunov–Schmidt reduction to reduce the equation $\iota(\lambda) = 0$ in a neighborhood of λ_* to an equation $\iota_{\mathrm{red}}(\lambda) = 0$, where

$$\iota_{\mathrm{red}}(\lambda) : \ \mathrm{N}(\iota(\lambda_*)) \longrightarrow \mathrm{Rg}(\iota(\lambda_*))^\perp$$

for λ close to λ_*. Nontrivial intersections are then given by zeros of the reduced Evans function

$$\mathcal{E}(\lambda) = \det(\iota_{\mathrm{red}}(\lambda)).$$

Recall that Floquet exponents λ and Floquet multipliers ρ of the linear period map Φ_{d} are related via $\rho = \exp(2\pi\lambda/\omega_{\mathrm{d}})$. We also denote by $i(\mathcal{A})$ the index of a Fredholm operator $\mathcal{A}$.

Lemma 4.1. *The linear operator $\Phi_{\mathrm{d}} - \rho$ on $L^2(\mathbb{R}, \mathbb{C}^n)$ is Fredholm if and only if $\lambda \notin \Sigma_{\mathrm{wt}}^- \cup \Sigma_{\mathrm{wt}}^+$. Furthermore, if $\lambda \notin \Sigma_{\mathrm{wt}}^- \cup \Sigma_{\mathrm{wt}}^+$, then the Fredholm index of $\Phi_{\mathrm{d}} - \rho$ on $L^2(\mathbb{R}, \mathbb{C}^n)$ is given by*

$$i(\Phi_{\mathrm{d}} - \rho) = i(\iota(\lambda)) = i_{\mathrm{wt}}^-(\lambda) - i_{\mathrm{wt}}^+(\lambda),$$

where $i_{\mathrm{wt}}^\pm(\lambda)$ are the relative Morse indices of the asymptotic wave trains defined in section 3.4. Last, if the Fredholm index is zero, then roots λ of the reduced Evans function $\mathcal{E}(\lambda)$ correspond to isolated eigenvalues ρ of Φ_{d}, and the order of a root λ is equal to the algebraic multiplicity of the corresponding Floquet multiplier ρ of Φ_{d}.

Proof. The relation between properties of the linearized period map Φ_{d} and the bundles $E_+^{\mathrm{s}}(\lambda)$ and $E_-^{\mathrm{u}}(\lambda)$ was shown in [49, Remark 2.5 and Theorem 2.6]. The fact that the order of roots of the reduced Evans function coincides with the algebraic multiplicity of the associated Floquet multiplier is a straightforward adaptation of the corresponding facts for eigenvalue problems of travelling waves. ■

Since we assumed that the wave trains are spectrally stable, we know that $\mathrm{Re}\,\lambda > 0$ lies in the resolvent set of Φ_{d} (in the Floquet-exponent space). This fact allows us to compute the Fredholm indices of the map

$$\iota_{\mathrm{d}}(\lambda) : \quad E_-^{\mathrm{cu}}(\lambda) \times E_+^{\mathrm{cs}}(\lambda) \longrightarrow Y, \qquad (\mathbf{u}^-, \mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+$$

for each of the four defect classes for λ close to zero, where we set $E_-^{\mathrm{cu}}(\lambda) := \mathrm{Rg}(\Phi_-^{\mathrm{cu}}(0,0))$ and $E_+^{\mathrm{cs}}(\lambda) := \mathrm{Rg}(\Phi_+^{\mathrm{cs}}(0,0))$. Note that, by using small exponential weights as outlined in the preceding section, we can choose the dichotomies Φ_-^{cu} and Φ_+^{cs} so that they depend analytically on λ for λ close to zero [49]. We then have the following result.

Lemma 4.2. *The Fredholm index i of $\iota_{\mathrm{d}}(0)$ is equal to*

$$\begin{aligned} i &= 2 && \textit{for sinks and contact defects,} \\ i &= 1 && \textit{for transmission defects,} \\ i &= 0 && \textit{for sources.} \end{aligned}$$

Proof. First, the Fredholm index of $\iota_\mathrm{d}(0)$ is given by the difference of the Morse indices of the projections $P^\mathrm{cu}_{\mathrm{wt},-}(0)$ and $P^\mathrm{cs}_{\mathrm{wt},+}(0)$ associated with the asymptotic wave trains. Indeed, we can apply Lemma 4.1 to the equation

$$\begin{aligned}\check{u}_\xi &= -\check{\eta}(\xi)\check{u} + \check{v},\\ \check{v}_\xi &= -\check{\eta}(\xi)\check{v} + D^{-1}[\omega_\mathrm{d}\partial_\tau\check{u} - c_\mathrm{d}\check{v} - f'(u_\mathrm{d}(\xi,\tau))\check{u} + \lambda\check{u}],\end{aligned}$$

where $\check{\eta}(\xi) = -\eta \operatorname{sign}\xi$ for some small $\eta > 0$. It therefore remains to compute the Morse indices.

For contact defects, the center subspace is two-dimensional, and the two spatial Floquet exponents $\nu_{1,2}$ are determined by (3.12) with $c_\mathrm{d} = c_\mathrm{g}$ so that $\nu_j^2 = \lambda/d_\| + \mathrm{O}(|\lambda|^{3/2})$, where $d_\| > 0$ by Hypothesis 3.2. In particular, ν_1 and ν_2 have opposite real parts for $\lambda > 0$. At $\lambda = 0$, the center-stable subspace $E^\mathrm{cs}_{\mathrm{wt},+}$ and the center-unstable subspace $E^\mathrm{cu}_{\mathrm{wt},-}$ are therefore both augmented by one-dimensional subspaces compared with the subspaces in the Fredholm index zero regime. Invoking the bordering Lemma 3.5 shows that the index of ι_d is two.

For the other defects, the center subspace is one-dimensional. To illustrate the idea, we consider the center-stable subspace $E^\mathrm{cs}_{\mathrm{wt},+}$ when $c_\mathrm{g}^+ < c_\mathrm{d}$, i.e., when transport occurs toward the defect. Lemma 3.6 shows that the real part of the critical Floquet exponent ν is positive for $\lambda > 0$, and the center subspace continues therefore as part of the center-stable subspace so that $E^\mathrm{cs}_{\mathrm{wt},+}(\lambda) = E^\mathrm{c}_{\mathrm{wt},+}(\lambda) \oplus E^\mathrm{s}_{\mathrm{wt},+}(\lambda)$. The same argument proves that $E^\mathrm{cs}_{\mathrm{wt},+}(\lambda) = E^\mathrm{s}_{\mathrm{wt},+}(\lambda)$ whenever $c_\mathrm{g}^+ > c_\mathrm{d}$ so that transport occurs away from the defect. The analogous statements for $E^\mathrm{cu}_{\mathrm{wt},-}$ show that the subspace $E^\mathrm{u}_{\mathrm{wt},-}$ is again augmented by a one-dimensional subspace when the transport is toward the defect so that $c_\mathrm{g}^- > c_\mathrm{d}$. Invoking the bordering Lemma 3.5, it is now straightforward to compute the Fredholm indices of ι_d for sources, sinks, and transmission defects. ■

4.3. Spectral stability of defects. We conclude this exposition of the linear theory by commenting on the effects of exponential weights on the spectra of defects. The discussion of weights in section 4.1 together with Lemma 4.1 can be used to infer useful properties of the spectra of the linearized period map in the exponentially weighted spaces

$$\begin{aligned}L^2_{\eta_-,\eta_+} &= \left\{u \in L^2_\mathrm{loc};\ \|u\|_{L^2_{\eta_-,\eta_+}} < \infty\right\},\\ \|u\|^2_{L^2_{\eta_-,\eta_+}} &= \int_{\mathbb{R}^-} |u(\xi)\mathrm{e}^{\eta_-\xi}|^2\,\mathrm{d}\xi + \int_{\mathbb{R}^+} |u(\xi)\mathrm{e}^{\eta_+\xi}|^2\,\mathrm{d}\xi.\end{aligned}$$

Indeed, the linearized period map $\Phi_\mathrm{d} - \rho$ is Fredholm on $L^2_{\eta_-,\eta_+}$ precisely if there are no spatial Floquet exponents $\nu_\pm$ of the asymptotic wave trains $\mathbf{u}_{\mathrm{wt},\pm}$ for which $\operatorname{Re}\nu_\pm = -\eta_\pm$. Invoking (3.12),

$$\lambda = \left[c_\mathrm{d} - c_\mathrm{g}^\pm\right]\nu + d_\|\nu^2 + \mathrm{O}(\nu^3)$$

with $d_\| > 0$, we see that the critical dispersion curve $\lambda(\nu)$ moves into the left half-plane provided the weights $\eta_\pm$ satisfy $\pm\eta_\pm > 0$ (thus enforcing localization of $u(\xi)$) when transport occurs toward the defect, whereas the weights $\eta_\pm$ have to satisfy $\pm\eta_\pm < 0$ (thus allowing exponential growth) when transport is away from the defect. In formulas, we need

$$\operatorname{sign}[c_\mathrm{d} - c_\mathrm{g}^\pm] = \operatorname{sign}\operatorname{Re}\nu \neq \operatorname{sign}[-\eta_\pm]$$

to ensure that the Floquet spectrum lies in $\operatorname{Re}\lambda < 0$, which is equivalent to choosing $\eta_\pm$ such that

$$\operatorname{sign}\eta_\pm = \operatorname{sign}[c_\mathrm{d} - c_\mathrm{g}^\pm].$$

In summary, we have proved the following result.

Lemma 4.3. *The essential Floquet spectrum of the period map, linearized about sinks, sources, and transmission defects, is contained in the open left half-plane when considered on $L^2_{\eta_-,\eta_+}$ for any choice of weights $\eta_\pm$ close to zero so that*

$$\begin{aligned} \eta_- < 0 < \eta_+ \qquad & \textit{for sinks with } c_\mathrm{g}^- > c_\mathrm{d} > c_\mathrm{g}^+, \\ \eta_-, \eta_+ > 0 \qquad & \textit{for transmission defects with } c_\mathrm{g}^\pm < c_\mathrm{d}, \\ \eta_- > 0 > \eta_+ \qquad & \textit{for sources with } c_\mathrm{g}^- < c_\mathrm{d} < c_\mathrm{g}^+. \end{aligned}$$

The essential Floquet spectrum of contact defects intersects the right half-plane in any $L^2_{\eta_-,\eta_+}$ space with exponential weights $\eta_- \neq 0$ or $\eta_+ \neq 0$.

The next lemma gives lower bounds for the multiplicity of $\lambda = 0$ as an isolated eigenvalue of the linearization Φ_d about each defect when considered on $L^2_{\eta_-,\eta_+}$ with $\eta_\pm$ chosen as in Lemma 4.3. We also refer to Figure 6.1.

Lemma 4.4. *Assume that there is a $\delta > 0$ such that*

$$\begin{aligned} \partial_\tau \mathbf{u}_\mathrm{d}(\xi, \cdot) &= -\mathbf{u}'_\mathrm{wt}(k_\pm\xi - \cdot) + \mathrm{O}(\mathrm{e}^{-\delta|\xi|}), \\ \partial_\xi \mathbf{u}_\mathrm{d}(\xi, \cdot) &= k_\pm \mathbf{u}'_\mathrm{wt}(k_\pm\xi - \cdot) + \mathrm{O}(\mathrm{e}^{-\delta|\xi|}) \end{aligned} \tag{4.7}$$

for $|\xi| \to \infty$. The geometric multiplicity of $\lambda = 0$ as an eigenvalue in the point spectrum of the linearized period map Φ_d posed on $L^2_{\eta_-,\eta_+}$ with $\eta_\pm$ chosen as in Lemma 4.3 is then at least equal to

$$\begin{aligned} 0 \quad & \textit{for sinks,} \\ 1 \quad & \textit{for transmission defects,} \\ 2 \quad & \textit{for sources.} \end{aligned}$$

Proof. Any linear combination of $\partial_\tau u_\mathrm{d}$ and $\partial_\xi u_\mathrm{d}$ satisfies the linearized equation and is time-periodic with the correct period. We have to check which of these linear combinations belongs to the space $L^2_{\eta_-,\eta_+}$ with $\eta_\pm$ chosen as in Lemma 4.3. For sinks, there is nothing to prove. For transmission defects with $c_\mathrm{d} > c_\mathrm{g}^\pm$, eigenfunctions need to be exponentially localized as $\xi \to \infty$. By assumption, $(k_+\partial_\tau + \partial_\xi)\mathbf{u}_\mathrm{d}(\xi, \cdot)$ generates a one-dimensional subspace of solutions that decay exponentially with rate δ at $\xi = \infty$. For sources, the exponential weights allow for exponential growth. Thus, any linear combination of $\partial_\tau u_\mathrm{d}$ and $\partial_\xi u_\mathrm{d}$ contributes to the null space of Φ_d. ■

The following result for contact defects has been proved in [50]. It will not be relevant for the robustness results in Theorem 1.4, although it is crucial for various dynamical stability considerations.

Theorem 4.5 (see [50]). *Let $\Omega = \{\lambda \in \mathbb{C};\ |\lambda| < \delta\}$, where $\delta > 0$ is sufficiently small. We consider a contact defect and assume that the null space of the map $\iota_\mathrm{d}(0)$ is two-dimensional.*

There exists then an analytic Evans function $\mathcal{E}(\lambda)$, defined for $\lambda \in \Omega \setminus \mathbb{R}^-$, whose roots, counted with their order, are in 1-1 *correspondence with Floquet multipliers* $\exp(2\pi\lambda/\omega_{\rm d})$, *counted with algebraic multiplicity, of the linearization $\Phi_{\rm d}$ of the period map about a contact defect. Moreover, $\mathcal{E}$ can be extended into $\lambda = 0$ as a C^1-function of $\sqrt{\lambda}$, and we have $\mathcal{E}(0) = 0$ and $\mathcal{E}'(0) \neq 0$ so that $\lambda = 0$ is a Floquet exponent with algebraic multiplicity one.*

Last, we state the following corollary which we shall exploit later. Note that the adjoint equation associated with (1.10) is given by

$$\omega_{\rm d} u_\tau = D u_{\xi\xi} - c_{\rm d} u_\xi + f'(u_{\rm d}(\xi,\tau))^* u. \tag{4.8}$$

We denote its period map by $\Phi_{\rm d}^{\rm ad}$.

Corollary 4.6. *Assume that $u_{\rm d}(\xi,\tau)$ is a transverse source. The null space of the adjoint operator $\Phi_{\rm d}^{\rm ad} - 1$ on $L^2(\mathbb{R},\mathbb{C}^n)$ is at least two-dimensional and contains two linearly independent functions $\psi_{\rm d}^{\rm c}(\xi,0)$ and $\psi_{\rm d}^{\omega}(\xi,0)$ that satisfy*

$$\int_{\mathbb{R}} \begin{pmatrix} \langle \psi_{\rm d}^{\rm c}(\xi,0), \partial_\xi u_{\rm d}(\xi,0)\rangle & \langle \psi_{\rm d}^{\rm c}(\xi,0), \partial_\tau u_{\rm d}(\xi,0)\rangle \\ \langle \psi_{\rm d}^{\omega}(\xi,0), \partial_\xi u_{\rm d}(\xi,0)\rangle & \langle \psi_{\rm d}^{\omega}(\xi,0), \partial_\tau u_{\rm d}(\xi,0)\rangle \end{pmatrix} {\rm d}\xi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Furthermore, the corresponding solutions $\psi_{\rm d}^{\rm c}(\xi,\tau)$ and $\psi_{\rm d}^{\omega}(\xi,\tau)$ of (4.8) *decay exponentially with a uniform rate as $\xi \to \pm\infty$ for all τ.*

Proof. The corollary is a consequence of [49, Lemma 5.1 and section 6] and Lemmas 4.3 and 4.4. Note that these results imply that $\Phi_{\rm d}^{\rm ad} - 1$ is bounded and Fredholm with index zero on $L^2_{-\eta}(\mathbb{R},\mathbb{C}^n)$, where $\eta = (\eta_-, \eta_+)$ is chosen as in Lemma 4.3. ■

5. Robustness of defects in oscillatory media.

5.1. Invariant manifolds. We begin by investigating the existence of stable and unstable manifolds for the ill-posed equation (4.2)

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_{\rm d}\partial_\tau u - c_{\rm d} v - f(u)]. \end{aligned}$$

Throughout this section, we fix an integer $1 \leq \ell < \infty$ and use the term *smooth* to refer to functions of class C^ℓ.

We define the *stable manifold* of the wave train $\mathbf{u}_{\rm wt}$ to be the set of initial conditions $\mathbf{u}_0$ for which there exist a solution $\mathbf{u}(\xi)$ of (4.2) with $\mathbf{u}(0) = \mathbf{u}_0$ and a continuous phase function $\theta(\xi)$ such that

$$\|\mathbf{u}(\xi) - \mathbf{u}_{\rm wt}(k\xi + \theta(\xi); k)\|_Y \to 0$$

as $\xi \to \infty$. The *unstable manifold* is defined in the same way with the limit considered as $\xi \to -\infty$. A *center manifold* $\mathcal{W}^{\rm c}$ is a locally invariant manifold that contains all solutions which stay in a sufficiently small neighborhood of the orbit $\mathbf{u}_{\rm wt}(k\xi - \cdot)$ for all $\xi \in \mathbb{R}$. Local invariance means that, for each $\mathbf{u}_0 \in \mathcal{W}^{\rm c}$, there exist a constant $\delta > 0$ and a solution $\mathbf{u}(\xi) \in \mathcal{W}^{\rm c}$, defined for $|\xi| < \delta$, with $\mathbf{u}(0) = \mathbf{u}_0$. Similarly, *center-stable* and *center-unstable manifolds* are locally invariant in the forward and backward ξ-directions, respectively, and contain all solutions that stay in a sufficiently small neighborhood of the wave-train orbit for all $\xi \in \mathbb{R}^+$

and $\mathbb{R}^-$, respectively. The *strong-stable manifold* of a point $\mathbf{u}_{\mathrm{wt}}(\theta)$ consists of all $\mathbf{u}_0$ for which there is a solution $\mathbf{u}(\xi)$ with $\|\mathbf{u}(\xi) - \mathbf{u}_{\mathrm{wt}}(k\xi + \theta - \cdot)\|_Y \to 0$ as $\xi \to \infty$. The *strong-unstable manifold* is defined analogously using the limit $\xi \to -\infty$.

We say that the above manifolds are *smooth* if they are smooth as manifolds *and* if the solutions $\mathbf{u}(\xi; \mathbf{u}_0)$ define smooth local semiflows for $\xi \geq 0$ and/or $\xi \leq 0$ for initial data $\mathbf{u}_0$ in these manifolds. We say that the center-stable manifold is smoothly fibered over the center manifold if it is the union of disjoint smooth fibers that intersect the center manifold transversely, depend continuously on the base point in the center manifold, and are mapped into each other by the local semiflow. We say that the above manifolds are *local* if they contain all solutions with the prescribed behavior whose initial data are close to the defect $\mathbf{u}_{\mathrm{d}}(0)$ and its τ-translates. The following result will be proved in Appendix A.1.

Theorem 5.1. *Each wave train has smooth local strong-stable, center-stable, center-unstable, and strong-unstable manifolds that depend smoothly on the parameters c and ω. If $c_{\mathrm{d}} \neq c_{\mathrm{g}}$, the center-stable and center-unstable manifolds are smoothly fibered by the union of the strong-stable and strong-unstable manifolds. The latter manifolds are parametrized by the asymptotic phase θ of the periodic wave train that the solutions approach. The center-stable (center-unstable) manifolds can be chosen so large that the given orbit $\{\mathbf{u}_{\mathrm{d}}(\xi);\ \xi \geq 0\}$ ($\{\mathbf{u}_{\mathrm{d}}(\xi);\ \xi \leq 0\}$) and its τ-translates are contained in their interior. The tangent space of the invariant manifolds evaluated at the defect $\mathbf{u}_{\mathrm{d}}(0)$ coincide with the corresponding ranges $E^j_\pm(0)$, with $j = \mathrm{ss}, \mathrm{cs}, \mathrm{s}, \mathrm{c}, \mathrm{u}, \mathrm{cu}, \mathrm{uu}$, of the exponential dichotomies for the linearized equation (4.3) with $\lambda = 0$ that we constructed in section 4.1.*

Corollary 5.2. *Sinks, sources, and transmission defects have an asymptotic phase. More precisely, there are constants $\delta > 0$ and $\theta_\pm \in \mathbb{R}$ such that*

$$\mathbf{u}_{\mathrm{d}}(\xi) = \mathbf{u}_{\mathrm{wt}}(k_\pm \xi - \cdot + \theta_\pm; k_\pm) + \mathrm{O}(\mathrm{e}^{-\delta|\xi|})$$

in Y as $|\xi| \to \infty$. The same estimate is true for the derivatives with respect to ξ and τ. In particular, hypothesis (4.7) in Lemma 4.4 is automatically satisfied for sinks, sources, and transmission defects.

The following proposition is a reformulation of Lemma 4.2 using Theorem 5.1.

Proposition 5.3. *Denote by E^{cu}_- and E^{cs}_+ the tangent spaces of the center-unstable and the center-stable manifold at $\mathbf{u}_{\mathrm{d}}(0)$. The injection map*

$$\iota: \quad W^{\mathrm{cu}}_- \times W^{\mathrm{cs}}_+ \longrightarrow Y, \qquad (\mathbf{u}^-, \mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+ \tag{5.1}$$

is Fredholm; i.e., it has continuous derivatives near $\mathbf{u}^- = \mathbf{u}^+ = \mathbf{u}_{\mathrm{d}}(0)$ which are Fredholm. Furthermore, the Fredholm index of the derivative ι' is given by

$$\begin{array}{lll} i = 2 & \text{if } c_{\mathrm{g}}^- \geq c_{\mathrm{d}} \geq c_{\mathrm{g}}^+ & (\textit{sinks and contact defects}), \\ i = 1 & \text{if } c_{\mathrm{g}}^\pm < c_{\mathrm{d}} \text{ or } c_{\mathrm{g}}^\pm > c_{\mathrm{d}} & (\textit{transmission defects}), \\ i = 0 & \text{if } c_{\mathrm{g}}^- < c_{\mathrm{d}} < c_{\mathrm{g}}^+ & (\textit{sources}). \end{array}$$

Last, $\partial_\tau \mathbf{u}_{\mathrm{d}}(0)$ and $\partial_\xi \mathbf{u}_{\mathrm{d}}(0)$ belong to $E^{\mathrm{cu}}_- \cap E^{\mathrm{cs}}_+$, and the null space of ι' is therefore at least two-dimensional.

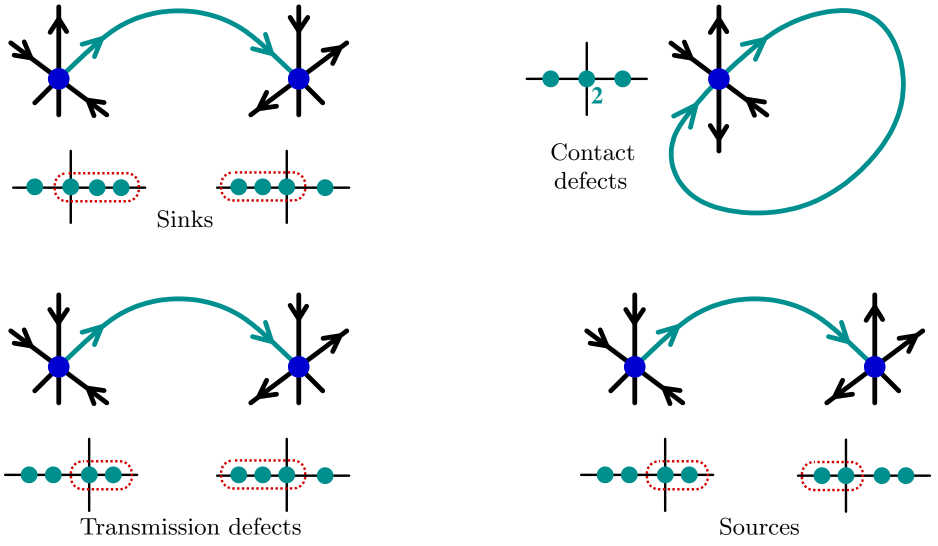


Figure 5.1. *Sinks, contact defects, transmission defects (with $c_{\mathrm{g}}^{\pm} < c_{\mathrm{d}}$), and sources correspond to homoclinic and heteroclinic orbits of the spatial-dynamical system* (4.2)*. The asymptotic equilibria symbolize periodic orbits, which in turn correspond to wave trains, after factoring the time-shift symmetry* (3.17)*. The insets show the spatial spectra of the linearization of* (4.2) *about the wave trains.*

The map ι can be constructed for all (ω, c) close to $(\omega_{\mathrm{d}}, c_{\mathrm{d}})$. If we view $W_{+}^{\mathrm{cs}}(\omega, c)$ as the graph of a smooth function G_{+}^{cs} that maps $W_{+}^{\mathrm{cs}}(\omega_{\mathrm{d}}, c_{\mathrm{d}})$ into E_{+}^{uu} and that depends smoothly on (ω, c), and if we define in an analogous fashion a function G_{-}^{cu}, then we can define a parameter-dependent injection map ι_{p}:

$$\begin{aligned}\iota_{\mathrm{p}}: \quad & W_{-}^{\mathrm{cu}}(\omega_{\mathrm{d}}, c_{\mathrm{d}}) \times W_{+}^{\mathrm{cs}}(\omega_{\mathrm{d}}, c_{\mathrm{d}}) \times \mathbb{R} \times \mathbb{R} \longrightarrow Y, \\ & \qquad (\mathbf{u}^{-}, \mathbf{u}^{+}, \omega, c) \longmapsto \mathbf{u}^{-} + G_{-}^{\mathrm{cu}}(\mathbf{u}^{-}, \omega, c) - \mathbf{u}^{+} - G_{+}^{\mathrm{cs}}(\mathbf{u}^{+}, \omega, c).\end{aligned}$$

Our goal is to show the persistence of defects upon varying the asymptotic wavenumbers and, possibly, external parameters. The geometric intuition that leads to our results is illustrated in Figure 5.1.

5.2. Sinks, sources, and transmission defects. For sinks, we set

$$\iota_{\mathrm{si}} = \iota_{\mathrm{p}}(\cdot, \cdot, \omega_{\mathrm{d}}, c_{\mathrm{d}}): \quad W_{-}^{\mathrm{cu}} \times W_{+}^{\mathrm{cs}} \longrightarrow Y. \tag{5.2}$$

For transmission defects with $c_{\mathrm{g}}^{\pm} < c_{\mathrm{d}}$, we fix k_{+}, which, on account of (1.7), implies $\omega = \omega_{\mathrm{nl}}(k_{+}) - ck_{+}$. We then define

$$\iota_{\mathrm{tr}}: \quad W_{-}^{\mathrm{cu}} \times W_{+}^{\mathrm{cs}} \times \mathbb{R} \longrightarrow Y, \qquad \iota_{\mathrm{tr}}(\mathbf{u}^{-}, \mathbf{u}^{+}, c) := \iota_{\mathrm{p}}(\mathbf{u}^{-}, \mathbf{u}^{+}, \omega_{\mathrm{nl}}(k_{+}) - ck_{+}, c). \tag{5.3}$$

Last, for sources, we set

$$\iota_{\mathrm{so}} = \iota_{\mathrm{p}}: \quad W_{-}^{\mathrm{cu}} \times W_{+}^{\mathrm{cs}} \times \mathbb{R} \times \mathbb{R} \longrightarrow Y. \tag{5.4}$$

Lemma 5.4. *The maps* ι_j *with* $j = \mathrm{si}, \mathrm{tr}, \mathrm{so}$ *are Fredholm maps with index* $i(\iota_j') = 2$. *If* ι_j' *is onto, then each defect is robust. In particular, sinks occur then as two-parameter families (parametrized by* (k_-, k_+)*), transmission defects as one-parameter families (parametrized by* k_+ *if* $c_\mathrm{g}^\pm < c_\mathrm{d}$*), and sources are isolated.*

Proof. The index formula follows from the bordering Lemma 3.5. Thus, suppose that ι_j' is onto. For sources, there is then a locally unique root of ι_so. For sinks, we can solve the equation $\iota_\mathrm{si} = 0$ for $(\mathbf{u}^-, \mathbf{u}^+)$, which lives in a complement of the two-dimensional null space of ι_si', as a function of $(\omega_\mathrm{d}, c_\mathrm{d})$ using the implicit function theorem. The result is a two-parameter family of sinks parametrized by $(\omega_\mathrm{d}, c_\mathrm{d})$. Note that $(\omega_\mathrm{d}, c_\mathrm{d})$ depend on the asymptotic wavenumbers (k_-, k_+) through (1.7) and (1.8):

$$(k_+, k_-) \longmapsto (\omega_\mathrm{d}, c_\mathrm{d}) = \left(\frac{\omega_\mathrm{nl}(k_-)k_+ - \omega_\mathrm{nl}(k_+)k_-}{k_+ - k_-}, \frac{\omega_\mathrm{nl}(k_+) - \omega_\mathrm{nl}(k_-)}{k_+ - k_-} \right).$$

Since the determinant of the derivative of this map is given by

$$\frac{(c_\mathrm{g}^+ - c_\mathrm{d})(c_\mathrm{g}^- - c_\mathrm{d})}{k_+ - k_-} \neq 0,$$

we can alternatively parametrize sinks by the asymptotic wavenumbers (k_-, k_+). Recall here that we assumed that sinks have $k_- \neq k_+$. Last, for transmission defects, the same arguments show that they persist as one-parameter families which are parametrized by the wavenumber k_+.

Note that the smooth dependence of $\mathbf{u}_\mathrm{d}(0)$ on parameters implies the smooth dependence of $\mathbf{u}_\mathrm{d}(\xi)$ on parameters on any finite interval $\xi \in [-L, L]$. For large values of ξ, the defects select a strong-stable fiber that itself depends smoothly on parameters and on its base point in a uniform topology, possibly after reparametrizing ξ. We refer to Appendix A.1 for the existence of these fibers. ■

We remark that we may also include external parameters, for instance, parameters in the nonlinearity f or the diffusion matrix D, in the definition of our maps ι_j. As a result, the defects considered in Lemma 5.4 depend smoothly on external parameters provided the maps ι_j' are onto.

The following proposition links spectral properties of transverse defects to the codimension of the range of ι_j'.

Proposition 5.5. *For sinks, sources, and transmission defects, the linear operator* ι_j' *with* $j = \mathrm{si}, \mathrm{tr}, \mathrm{so}$ *is onto if and only if the defect is transverse, i.e., if and only if the spectrum of the defect at the origin is minimal.*

Proof. We have to show that, under the spectral assumptions for elementary transverse defects, the map ι_j is onto if and only if the critical spectrum is minimal. This then proves the robustness theorem for sinks, transmission defects, and sources, invoking Lemma 5.4.

Sinks. Suppose that ι_si' is onto. Proposition 5.3 and Lemma 5.4 imply that each bounded solution to (4.3) with $\lambda = 0$,

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f'(u_\mathrm{d}(\xi, \cdot))u], \end{aligned} \tag{5.5}$$

is a linear combination of $\partial_\tau \mathbf{u}_\mathrm{d}$ and $\partial_\xi \mathbf{u}_\mathrm{d}$. We claim that $E^\mathrm{ss}_+(0)$ and $E^\mathrm{uu}_-(0)$ intersect trivially. Indeed, as a consequence of Corollary 5.2, the linear combinations $(k_+\partial_\tau + \partial_\xi)\mathbf{u}_\mathrm{d}(\xi,\cdot)$ and $(k_-\partial_\tau + \partial_\xi)\mathbf{u}_\mathrm{d}(\xi,\cdot)$ are the unique linear combinations that result in exponentially decaying functions as $\xi \to \infty$ and $\xi \to -\infty$, respectively. Since $k_- \neq k_+$, however, these combinations do not match, and none of the linear combinations will therefore decay at both $\xi = \infty$ and $\xi = -\infty$. This shows that there are no solutions to (5.5) that decay exponentially as $|\xi| \to \infty$, and the sink is therefore transverse.

Conversely, assume that the dimension of the null space of ι'_si is at least three; then there exists a solution $\mathbf{u}_*$ to (5.5) which is linearly independent of $\partial_\xi \mathbf{u}_\mathrm{d}$ and $\partial_\tau \mathbf{u}_\mathrm{d}$. Since the spatial Floquet exponent $\nu = 0$ of the equation for the asymptotic wave trains is simple, we may write the solution $\mathbf{u}_*$ as

$$\mathbf{u}_*(\xi,\cdot) = a_\pm \mathbf{u}'_\mathrm{wt}(k_\pm\xi - \cdot; k_\pm) + \mathrm{O}(\mathrm{e}^{-\delta|\xi|})$$

for some $\delta > 0$ as $\xi \to \pm\infty$. Exploiting the expansion for $\mathbf{u}_\mathrm{d}$ and the fact that $k_- \neq k_+$, we find constants c_1 and c_2 so that

$$\mathbf{u}_* + c_1\partial_\tau \mathbf{u}_\mathrm{d} + c_2 \partial_\xi \mathbf{u}_\mathrm{d}$$

decays exponentially as $|\xi| \to \infty$. This shows that the null space of the linearized period map, considered in $L^2_{\eta_-,\eta_+}$ with $\eta_- > 0 > \eta_+$ close to zero, is not trivial.

Sources. Assume that ι'_so is onto. The first components of the two bounded, linearly independent solutions $\partial_\xi \mathbf{u}_\mathrm{d}$ and $\partial_\tau \mathbf{u}_\mathrm{d}$ of (5.5) form a two-dimensional subspace in the null space of the linearized period map Φ_d considered in $L^2_{\eta_-,\eta_+}$ with $\eta_- < 0 < \eta_+$. Since the null space of ι'_so contains all solutions with sufficiently mild exponential growth, the geometric multiplicity of $\lambda = 0$ is equal to two. We need to exclude generalized eigenfunctions. Assume therefore that $\check{u}_*$ satisfies

$$\check{u}_*(\cdot, 2\pi) = \Phi_\mathrm{d}\check{u}_*(\cdot,0) = \check{u}_*(\cdot,0) + (c_1\partial_\tau + c_2\partial_\xi)u_\mathrm{d}(\cdot,2\pi).$$

If we set

$$u_*(\xi,\tau) = \check{u}_*(\xi,\tau) - \frac{\tau}{2\pi}[c_1\partial_\tau + c_2\partial_\xi]u_\mathrm{d}(\xi,\tau),$$

then $\mathbf{u}_* = (u_*, \partial_\xi u_*)$ is in Y for all ξ and satisfies

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f'(u_\mathrm{d}(\xi,\cdot))u - (\omega_\mathrm{d}/2\pi)(c_1\partial_\tau u_\mathrm{d}(\xi,\cdot) + c_2 v_\mathrm{d}(\xi,\cdot))], \end{aligned} \tag{5.6}$$

where $\mathbf{u}_\mathrm{d} = (u_\mathrm{d}, v_\mathrm{d})$. Equation (5.6) is exactly the variational equation in the extended phase space, where ω and c are considered as additional variables. In particular, this equation in the extended phase space has a bounded solution that is linearly independent of the derivatives of the defect with respect to τ and ξ. As a consequence, the null space of ι'_so is at least three-dimensional, and ι'_so cannot be onto. This proves that the algebraic multiplicity of $\lambda = 0$ is equal to two.

Conversely, assume that the algebraic multiplicity of the linearized period map $\Phi_{\rm d}$ is two but that $\iota'_{\rm so}$ is not onto. By Fredholm theory, either the null space of $\partial_{(\mathbf{u}^-,\mathbf{u}^+)}\iota_{\rm so}$ has dimension larger than two or else the ranges of $\partial_{(\omega,c)}\iota_{\rm so}$ and of $\partial_{(\mathbf{u}^-,\mathbf{u}^+)}\iota_{\rm so}$ have a nontrivial intersection. In the first case, we can easily construct additional eigenfunctions and, in the second case, generalized eigenfunctions of the linearized period map $\Phi_{\rm d}$, contradicting our starting assumption.

Transmission defects. We assume that $\iota'_{\rm tr}$ is onto. Arguing as for sources, we conclude that the geometric multiplicity of $\lambda = 0$ is equal to one with the null space of $\Phi_{\rm d}$ spanned by $(k_+\partial_\tau + \partial_\xi)u_{\rm d}$. We argue by contradiction and assume there is a generalized eigenvector. As for sources, this assumption gives a solution of

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_{\rm d}\partial_\tau u - c_{\rm d} v - f'(u_{\rm d}(\xi,\tau))u - v_{\rm d}(\xi,\tau) - k_+\partial_\tau u_{\rm d}(\xi,\tau)], \end{aligned} \tag{5.7}$$

which is bounded as $\xi \to -\infty$ and decays exponentially as $\xi \to \infty$. The construction of $\iota_{\rm tr}$ in (5.3) shows that (5.7) is the variational equation associated with $\partial_c\iota$. Therefore, $\iota'_{\rm tr}$ will have a null space of dimension larger than two, contradicting our assumption. The converse follows similarly. ■

5.3. Contact defects. Recall that $E^{\rm c}_\pm$ is two-dimensional for contact defects. We shall need the two injection maps

$$\begin{aligned} \iota_+ :&\quad W^{\rm cu}_- \times W^{\rm ss}_+ \longrightarrow Y, \quad (\mathbf{u}^-,\mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+, \\ \iota_- :&\quad W^{\rm uu}_- \times W^{\rm cs}_+ \longrightarrow Y, \quad (\mathbf{u}^-,\mathbf{u}^+) \longmapsto \mathbf{u}^- - \mathbf{u}^+. \end{aligned} \tag{5.8}$$

Both maps depend smoothly on additional parameters if any are present. Counting dimensions, the bordering Lemma 3.5 shows that $\iota_\pm$ are both Fredholm maps with index zero.

Lemma 5.6. *Contact defects occur as robust one-parameter families in the asymptotic wavenumber $k = k_\pm$ provided both ι'_+ and ι'_- are onto (and therefore invertible).*

Proof. First note that, since ι_+ is onto, so is the map ι from (5.1). In particular, via Lyapunov–Schmidt reduction, the intersection of center-unstable and center-stable manifolds persists. We claim that the defect lies neither in the strong-stable nor the strong-unstable manifold. Indeed, if it did, ι_+ or ι_- would have a nontrivial kernel and could therefore not be onto. The complement of the strong-stable manifold in the center-stable manifold of the wave train is an open subset of the center-stable manifold (initially in a neighborhood of the wave train, but then also in a neighborhood of $\mathbf{u}_{\rm d}(0)$ upon using continuity of the evolution on the center-stable manifold). In particular, the perturbed intersection points in $W^{\rm cs}_+$ still belong to the interior of the center-stable manifold. A similar description is true for the center-unstable manifold which shows persistence for nearby wavenumbers and parameters upon varying $c_{\rm d} = \omega'_{\rm nl}(k)$ appropriately. Smooth dependence of contact defects on parameters in a ξ-uniform topology is achieved after an appropriate reparametrization of the spatial time variable ξ. ■

Proposition 5.7. *The injection maps $\iota'_\pm$ are both onto if and only if the minimal-spectrum assumption holds, i.e., if and only if the contact defect is transverse.*

Proof. The null spaces of ι'_- and ι'_+ consist exactly of eigenfunctions of the linearized period map $\Phi_{\rm d}$ in $L^2_{\eta_-,\eta_+}$ with exponential weights $\eta_- = \eta_+ < 0$ and $\eta_- = \eta_+ > 0$, respectively. ■

Lemma 5.6 and Proposition 5.7 implicitly show that higher-dimensional intersections of center-stable and center-unstable manifolds would contribute to the null spaces of ι_+ and ι_-. They would, however, typically not contribute an additional root of the extension into $\lambda = 0$ of the Evans function $\mathcal{E}(\lambda)$ of Theorem 4.5. Indeed, we proved in [50] that roots of the extended Evans function are generated by solutions of the linearized equation that decay like $1/|\xi|$. The additional solutions $\mathbf{u}_*$ of the linearized equation, which correspond to the additional direction in the intersection of center-stable and center-unstable manifolds, would have asymptotics of the form $\mathbf{u}_* \approx a_\pm \mathbf{u}'_{\mathrm{wt}}$ as $\xi \to \pm\infty$. We expect $a_+ \neq a_-$, and since $\partial_\tau u_{\mathrm{d}}$ and $\partial_\xi u_{\mathrm{d}}$ have the same asymptotics at $\xi = \infty$ and at $\xi = -\infty$ (recall that $k_- = k_+$ for contact defects), we cannot make $a_- = a_+ = 0$ by adding appropriate linear combinations of $\partial_\tau \mathbf{u}_{\mathrm{d}}$ and $\partial_\xi \mathbf{u}_{\mathrm{d}}$.

Similarly, the orbit-flip situation where the contact defect lies in the strong-stable or the strong-unstable manifold cannot be excluded from considerations of the extended Evans function that we constructed in [50]. We refer to section 6.4 for some puzzling consequences of these facts.

5.4. Proof of Theorem 1.4. For transverse sinks, sources, and transmission defects, Proposition 5.5 shows that we can apply Lemma 5.4 to find locally unique and robust families of defects. On the other hand, the maps ι'_j, evaluated along this family, remain onto, which, using the equivalence proved in Proposition 5.5, shows that the defects in the family are transverse, i.e., that they satisfy the minimal-spectrum assumption. For contact defects, the same conclusion is reached by combining Lemma 5.6 and Proposition 5.7.

6. Stability, bifurcations, pinning, and truncation. We collect various consequences of our results and point out a number of open problems related to the proposed classification of defects in one-dimensional media.

6.1. Stability. An important feature of transverse defects are the point and essential spectra of the linearized period map. The following theorem summarizes our findings, illustrated also in Figure 6.1, for the temporal Floquet exponents of Φ_{d}.

Theorem 6.1. *The following is true for transverse sinks, contact defects, transmission defects, and sources. Choose weights $\eta = (\eta_-, \eta_+)$ close to zero such that*

$$\begin{aligned} &\eta_- < 0 < \eta_+ && \textit{for sinks,} \\ &\eta_- = \eta_+ = 0 && \textit{for contact defects,} \\ &\eta_-, \eta_+ > 0 && \textit{for transmission defects with } c_{\mathrm{g}}^\pm < c_{\mathrm{d}}, \\ &\eta_- > 0 > \eta_+ && \textit{for sources.} \end{aligned} \tag{6.1}$$

The Floquet spectra, in a sufficiently small neighborhood of $\lambda = 0$, of the period map Φ_{d} posed on $L^2_\eta(\mathbb{R}, \mathbb{C}^n)$ are as shown in Figure 6.1.

Transverse sinks that have no additional isolated eigenvalues in the closed right half-plane are nonlinearly asymptotically stable for (1.1) in that perturbations decay in L^2_η for weights η as in (6.1). For transverse sources, the two eigenfunctions of the adjoint operator $\Phi_{\mathrm{d}}^{\mathrm{ad}} - 1$ on $L^2_{-\eta}(\mathbb{R}, \mathbb{C}^n)$ are exponentially localized.

Proof. We proved in Lemma 4.3 that the essential Floquet spectrum of sinks, sources, and transmission defects lies in the left half-plane for the weights chosen in (6.1). For contact

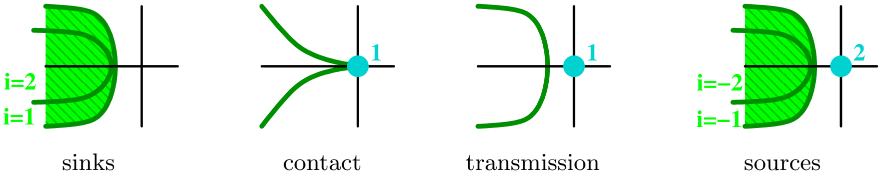


Figure 6.1. *Plotted are the Floquet spectra near $\lambda = 0$ of the maps Φ_d, posed on $L^2_\eta(\mathbb{R}, \mathbb{C}^n)$, for transverse sinks, contact defects, transmission defects, and sources with weights $\eta = (\eta_-, \eta_+)$ as in (6.1). Solid lines denote the Floquet spectra of the asymptotic wave trains, while the shaded areas correspond to regions where $\Phi_\mathrm{d} - \rho$ is Fredholm with nonzero index i. The bullets label eigenvalues with the attached number indicating their multiplicity (the algebraic and geometric multiplicities coincide).*

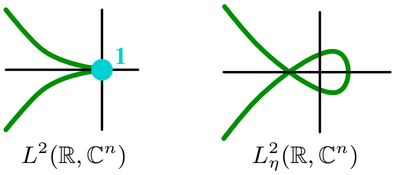

1
$L^2(\mathbb{R}, \mathbb{C}^n)$
$L^2_\eta(\mathbb{R}, \mathbb{C}^n)$

Figure 6.2. *Plotted are the Floquet spectra of contact defects near $\lambda = 0$ of the maps Φ_d on $L^2(\mathbb{R}, \mathbb{C}^n)$ and $L^2_\eta(\mathbb{R}, \mathbb{C}^n)$ for nonzero $\eta_- = \eta_+$ close to zero. The map $\Phi_\mathrm{d} - \rho$ is Fredholm with index zero off the solid lines.*

defects, there is an additional subtlety: The curve that corresponds to the linear dispersion relation of the asymptotic wave train has a cusp at $\lambda = 0$. Indeed, the dispersion relation in the frame moving with speed c_d is given by (3.12)

$$\lambda_\mathrm{lin}(\nu) = [c_\mathrm{d} - c_\mathrm{g}]\nu + d_\|\nu^2 + d_3\nu^3 + \mathrm{O}(\nu^4),$$

where the coefficients $d_\|$ and d_3 are real. Since we have $c_\mathrm{d} = c_\mathrm{g}$, we obtain

$$\lambda(\mathrm{i}\gamma) = -d_\|\gamma^2 + \mathrm{i}d_3\gamma^3 + \mathrm{O}(\gamma^4) \tag{6.2}$$

so that $\mathrm{Re}\,\lambda = [d_3\,\mathrm{Im}\,\lambda/d_\|]^{2/3}$ has a cusp point at the origin as claimed provided $d_3 \neq 0$.

Lemmas 3.6 and 4.1 together show that the Fredholm index for sinks and sources jumps by $+1$ and -1, respectively, when λ crosses Σ^-_wt or Σ^+_wt from right to left, while the Fredholm index of contact and transmission defects is zero to the left of the Floquet spectra $\Sigma^-_\mathrm{wt} \cup \Sigma^+_\mathrm{wt}$. The statements about the algebraic and geometric multiplicity of eigenvalues at $\lambda = 0$ follow at once from Lemma 4.4 and the proof of Proposition 5.5. Theorem 4.5 shows that the Evans function of transverse contact defects has a simple root at $\lambda = 0$.

Nonlinear stability of spectrally stable sinks in $L^2_\eta(\mathbb{R}, \mathbb{C}^n)$ follows easily since the period map is a contraction, while the nonlinearity f is well defined and smooth on $L^2_\eta(\mathbb{R}, \mathbb{C}^n)$ provided the weights are chosen as in (6.1). We refer to [11, 54] for details. Last, the statement about the exponential localization of the two adjoint eigenfunctions for sources is simply Corollary 4.6. ■

The spectrum of contact defects in L^2_η is shown in Figure 6.2. In particular, it is an immediate consequence of (6.2) that the essential spectrum is always unstable whenever the rates $\eta_- = \eta_+$ of the exponential weights are not zero.

The above theorem implies, in particular, that spectrally stable transverse sinks are nonlinearly asymptotically stable under small exponentially localized perturbations;[7] these perturbations decay exponentially in time, and we recover the *same* sink with no shift in its spatial or phase position since there is no point spectrum in the closed right half-plane.

The nonlinear stability of a specific transmission defect has been proved recently in a context of small-amplitude background waves [17]. For contact defects, the only related results we know of are in the context of conservation laws where Howard [25] recently investigated degenerate shock waves. We are not aware of any nonlinear stability results for sources. We expect, however, results along the following lines.

Perturbations of transmission defects should relax to a spatio-temporal translate of the original defect that preserves the relative phase of the wave train whose group velocity is directed toward the defect. Thus, if $c_\mathrm{g}^+ < c_\mathrm{d}$, say, then an initial condition $u(x,0)$ close to the transmission defect $u_\mathrm{d}(x,0)$ evolves toward a shifted transmission defect

$$u(x,t) \longrightarrow u_\mathrm{d}(\check{x} - c_\mathrm{d}\check{t}, \check{t})$$

as $t \to \infty$, where $k\check{x} - \omega\check{t} = kx - \omega t$. This is immediately clear on the linearized level, using the exponential weights (6.1), and has been proved for the example considered in [17].

Sources, on the other hand, should have a unique position and a unique phase as evidenced by the fact that the adjoint eigenfunctions are exponentially localized. This implies that the spectral projection onto the two-dimensional eigenspace consisting of space and time translations of the defect is well defined in L^2_η with weights as in (6.1). Thus, the linear analysis predicts that the shifts in space $\check{x} - x$ and time $\check{t} - t$ are uncorrelated.

6.2. Phase matching. Phase matching at defects has recently been discussed quite extensively in the physics literature [23, 1, 36, 37]. To explain this issue, we fix a continuous function $\vartheta(k)$ and define the phase of a wave train to be its argument

$$\phi[u_\mathrm{wt}(\phi;k)] := \phi - \vartheta(k) \tag{6.3}$$

after subtracting the fixed phase shift $\vartheta(k)$. Note that it is natural to choose $\vartheta(k)$ subject to $\vartheta(k) = -\vartheta(-k)$ to account for the reflection symmetry of the reaction-diffusion system (1.1).

Elementary defects $\check{u}_\mathrm{d}(x,t)$ with asymptotic phase, i.e., sinks, transmission defects, and sources, satisfy

$$\check{u}_\mathrm{d}(x,t) - u_\mathrm{wt}^\pm(k_\pm x - \omega_\mathrm{nl}(k_\pm)t + \theta_\pm; k_\pm) \longrightarrow 0$$

as $x \to \pm\infty$ for appropriate constant phase corrections $\theta_\pm$. If we wish to measure the phase mismatch $[\phi]$ across the characteristic $x - c_\mathrm{d}t = \xi_\mathrm{pm}$ for some fixed $\xi_\mathrm{pm} \in \mathbb{R}$, we can do so by defining the two phases

$$\begin{aligned}\phi_\pm(t) &:= \phi[u_\mathrm{wt}^\pm(k_\pm x - \omega_\mathrm{nl}(k_\pm)t + \theta_\pm; k_\pm)] = \phi[u_\mathrm{wt}^\pm(k_\pm\xi_\mathrm{pm} - \omega_\mathrm{d}t + \theta_\pm; k_\pm)] \\ &= k_\pm\xi_\mathrm{pm} - \omega_\mathrm{d}t + \theta_\pm - \vartheta(k_\pm)\end{aligned}$$

[7] Nonlinear stability results using weaker polynomial weights have been established in specific examples [29, 15].

and then take their difference to get

$$[\phi] := \phi_+(t) - \phi_-(t) = [k]\xi_{\text{pm}} + [\theta] - [\vartheta], \tag{6.4}$$

where we use the notation $[F] = F(k_+) - F(k_-)$ to denote jumps of functions across the defect. We call $[\phi]$ the *phase slip* across the defect along the characteristic line $x - c_{\text{d}}t = \xi_{\text{pm}}$. Note that the position ξ_{pm} along which we measure the phase slip is somewhat arbitrary. We also emphasize that the definition of the phase slip relies on the assumption that the frequencies of the asymptotic wave trains coincide when computed in the frame that moves with the speed c_{d} of the defect.

The phase slip $[\phi]$ for transmission defects with $k_- = k_+$ is given by $[\phi] = [\theta] = \theta_+ - \theta_-$ and therefore is independent of the position $\xi_\pm$. Instead, it depends only on the difference between the asymptotic phase corrections $\theta_\pm$. For contact defects, we can also define a phase slip. Contact defects satisfy [11, 50]

$$\check{u}_{\text{d}}(x,t) - u_{\text{wt}}(kx - \omega_{\text{nl}}(k)t + \theta_\pm + \check{\theta}\log(x - c_{\text{d}}t); k) \longrightarrow 0$$

so that their phase slip is again given by $[\phi] = [\theta]$ since the logarithmic terms cancel.

In summary, sinks and sources have a phase slip $[\phi] = [k]\xi_{\text{pm}} + [\theta - \vartheta]$ that is periodic in ξ_{pm} with period $2\pi/[k]$. In particular, we can always arrange to obtain the zero phase slip $[\phi] = 0$ by measuring at the "correct" characteristic line. For contact defects and for transmission defects with $[k] = 0$, the phase slip $[\phi] = [\theta]$ is an intrinsic property of the defect that, in particular, does not depend on where we measure it.

We may use the phase slip to track the *position* $x = \xi_{\text{pm}} + c_{\text{d}}t$ of sinks and sources (even though the phase slip defines the position only in S^1 which is a minor complication when we consider small perturbations of a given defect). A given phase slip $[\phi]$ therefore gives a certain well-defined position $x = \xi_{\text{pm}} + c_{\text{d}}t$ of the defect. We may then compare the position of an unperturbed defect with the position of the defect after adding a small perturbation. For sinks, the asymptotic relaxation of perturbations to exactly the same sink guarantees that the position of the perturbed defect is the same as the position of the unperturbed defect, whereas we expect shifts of the position for sources.

6.3. Reflection symmetries.

The reaction-diffusion system (1.1) respects the reflection symmetry $x \mapsto -x$ in addition to the spatio-temporal translation symmetries. Thus, defects with speed zero are somewhat distinguished. We therefore set $c = 0$ in this section so that x is the spatial variable in the modulated-wave equation (4.2)

$$\begin{aligned} u_x &= v, \\ v_x &= D^{-1}[\omega_{\text{d}}\partial_\tau u - f(u)]. \end{aligned} \tag{6.5}$$

The reflection symmetry $x \mapsto -x$ for (1.1) manifests itself as a reversibility for (6.5). In fact, exploiting also the time-shift symmetry of (6.5), both

$$\mathcal{R}_0 : (u,v)(\tau) \mapsto (u,-v)(\tau) \qquad \text{and} \qquad \mathcal{R}_\pi : (u,v)(\tau) \mapsto (u,-v)(\tau+\pi) \tag{6.6}$$

are reversers that anticommute with the right-hand side of (6.5).

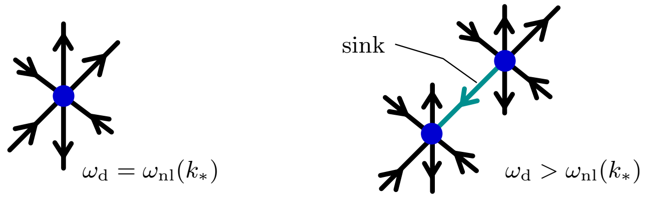

$\omega_\mathrm{d} = \omega_\mathrm{nl}(k_*)$
sink
$\omega_\mathrm{d} > \omega_\mathrm{nl}(k_*)$

Figure 6.3. *The unfolding of a saddle-node wave train generates a "small-amplitude" sink that connects two wave trains (the sketch assumes that $\omega''_\mathrm{nl}(k_*) > 0$).*

We may now seek symmetric defects to (1.1) which correspond to reversible homoclinic and heteroclinic orbits of (6.5). Reversible connecting orbits can be found as intersections of the center-unstable manifold of the asymptotic wave train at $x = -\infty$ with the fixed-point spaces $\mathrm{Fix}(\mathcal{R}_0) = \{(u, v);\ v = 0\}$ or $\mathrm{Fix}(\mathcal{R}_\pi) = \{(u, v);\ v(\tau) = -v(\tau + \pi)\}$ of the reversers $\mathcal{R}_0$ or $\mathcal{R}_\pi$, respectively. Note that the resulting defects have $k_+ = -k_-$, and therefore $c_\mathrm{g}^+ = -c_\mathrm{g}^-$ since $\omega_\mathrm{nl}(-k) = -\omega_\mathrm{nl}(k)$. In particular, transmission defects cannot be symmetric, since the group velocities to the left and right can have the same sign only if they both vanish.

Using the map

$$\iota_j: \quad W_-^\mathrm{cu} \times \mathrm{Fix}(\mathcal{R}_j) \longrightarrow Y, \quad (\mathbf{u}^-, \mathbf{u}^0) \longmapsto \mathbf{u}^- - \mathbf{u}^0$$

for $j = 0, \pi$, we can again compute its Fredholm index and compare it with robustness properties of symmetric defects. We obtain that symmetric sinks (with respect to either $\mathcal{R}_0$ or $\mathcal{R}_\pi$) arise as robust one-parameter families, while symmetric sources and contact defects are robust and occur for isolated values of the parameter k_- (note that $c = 0$ is required for symmetric defects). We remark that both $\mathcal{R}_\pi$-symmetric sources [38] and $\mathcal{R}_\pi$-symmetric contact defects [58] have been observed in experiments.

6.4. Bifurcations.

We address instabilities of defects and transitions between different defect types. From the spatial-dynamics perspective, this amounts to investigating homoclinic and heteroclinic bifurcations.

Saddle-node bifurcations of wave trains. As outlined in Example II of section 1.3, stable sinks with "small" amplitude are created at saddle-node bifurcations of wave trains [26, 11]. Indeed, if we consider the modulated-wave equation near a wave train in the frame that moves with its group velocity $c_\mathrm{d} = c_\mathrm{g}$, then the wave train has an algebraically double spatial Floquet exponent $\nu = 0$ (see section 3.4 and Figure 3.1). If we keep the defect speed fixed at $c_\mathrm{d} = c_\mathrm{g}(k_*)$, then the saddle node can be unfolded by varying the frequency ω_d near $\omega_\mathrm{nl}(k_*)$, where k_* is the wavenumber of the wave train we started with. The vector field on the resulting two-dimensional center manifold is invariant under the temporal time-shift symmetry and is given by

$$\phi_\xi = q + \mathrm{O}(q^2), \qquad \lambda''_\mathrm{lin}(0) q_\xi = \breve{\omega} - \omega''_\mathrm{nl}(k) q^2 + \mathrm{O}(|\breve{\omega}^2| + |\breve{\omega} q| + |q|^3), \tag{6.7}$$

where (ϕ, q) correspond roughly to phase and wavenumber, and where $\omega = \breve{\omega} + \omega_\mathrm{nl}(k_*)$. To leading order, we therefore recover the steady-state equation of the Burgers equation (1.19), which, as discussed in section 1.3, admits stable sinks. We refer to Figure 6.3 for an illustration.

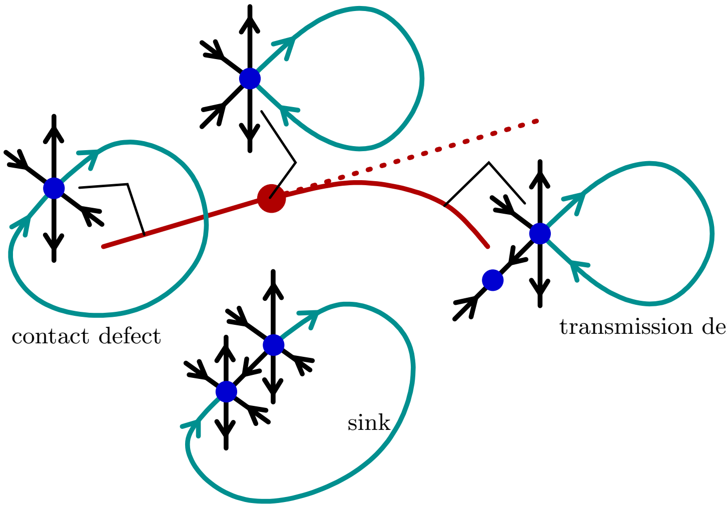


Figure 6.4. *The saddle-node homoclinic bifurcation in the two-dimensional parameter space.*

Large-amplitude sinks arise close to contact defects. If we vary frequency or speed so that the saddle-node wave train splits into two wave trains, then the saddle-node homoclinic orbit that corresponds to the contact defect becomes a heteroclinic orbit—in fact, a sink—that connects the two wave trains (as illustrated in the lower left half of Figure 6.4). It is an interesting problem to determine whether the resulting sink is stable or not. On account of Theorem 4.5, the Evans function of the contact defect will have a simple zero at the origin, while the sink does not have an eigenvalue at $\lambda = 0$. Thus, we expect that the sink will have either a weakly stable or a weakly unstable eigenvalue near zero.[8]

We remark that the bottom of Figure 6.4 also illustrates that the same wave trains can accommodate several distinct defects. Indeed, the two wave trains in the bottom plot of Figure 6.4 are connected by "small-amplitude" and "large-amplitude" sinks.

Folds. Next, we discuss what happens when the minimal-spectrum assumption, which is equivalent to the transversality condition for center-stable and center-unstable manifolds in the proof of Theorem 1.4, is violated. In this case, sinks acquire a simple eigenvalue at $\lambda = 0$ in the space L^2_η with $\operatorname{sign} \eta_\pm = \pm 1$. This degeneracy corresponds to a tangency of the center-unstable and center-stable manifolds of the wave trains in the modulated-wave equation (4.2). Thus, if the tangency is quadratic as expected, it will persist along a curve in (k_-, k_+)-parameter space. On one side of this curve, there exists a pair of sinks, while there are no sinks on its other side. It is not hard to prove that the additional critical eigenvalue near $\lambda = 0$ will stabilize one of the two sinks and destabilize the other one. The scenario for transmission defects and sources is similar since the linearization acquires a Jordan block of length two. For transmission defects, we then see two defects for $k < k_*$, say, and none for $k > k_*$, while folds for sources occur only when an additional external parameter is present.

Contact defects behave differently. If the center-stable and center-unstable manifolds

[8] **Note added in proof.** We have recently shown that this eigenvalue is, in fact, always stable and therefore lies in the open left half-plane.

intersect tangentially, we expect again a standard saddle-node bifurcation of contact defects in the modulated-wave equation that is unfolded by the wavenumber k. Since the asymptotic wavenumbers are identical, the tangency will, however, *not* generate a localized eigenfunction of the linearization. In particular, the Evans function near $\lambda = 0$ does not change at all. Thus, none of the two contact defects will acquire any additional eigenvalues, and both of them will be spectrally stable! Instead, an unstable root arises for the Evans functions associated with the maps $\iota_\pm$ from (5.8) which have the wrong Morse index. Of course, the two stable contact defects are not close to each other in the supremum norm, which precludes viewing one of them as a small perturbation of the other one.

Locking and unlocking via flip bifurcations. Contact defects are also destroyed at values of k at which the center-unstable manifold intersects the center-stable manifold along the strong-stable manifold of the asymptotic wave train. The associated homoclinic flip bifurcation has been analyzed in [6] for ODEs, and the resulting bifurcation diagram is shown in Figure 6.4. We remark that it is straightforward to obtain the same for (4.2) by using exponential dichotomies and foliations of center-stable and center-unstable manifolds. Note that the contact defect exists for wavenumbers below a critical wavenumber, say, and its speed is therefore determined solely by the group velocity. Above the critical wavenumber, the defect changes into a transmission defect whose speed is now determined by a Melnikov integral that depends on the profile of the defect. Thus, we may refer to this bifurcation as an unlocking bifurcation at which the speed of the defect unlocks from the group velocity of the underlying wave trains. We remark that neither the transmission defect nor the contact defect acquires any additional eigenvalues during this transition. Phenomenologically, the unlocking transition is preceded by an increasing localization of the defect structure. At the bifurcation point, the convergence toward the wave trains changes from algebraic to exponential. The unlocked transmission defect approaches the wave train ahead[9] of it at a uniform exponential rate, whereas the relaxation toward the wave train behind the defect occurs at a weak exponential rate. A similar phenomenon occurs when transmission defects bifurcate from pulses at parameter values where the homogeneous background undergoes a Turing instability [44].

A more dramatic unlocking bifurcation occurs if we allow an additional external parameter μ to vary. It can then happen that both the strong-stable and strong-unstable manifolds of a saddle-node periodic orbit intersect, which can be interpreted as the simultaneous unlocking of the contact defect at both end points. The analysis of this bifurcation is very similar to the one for localized inhomogeneities of wave trains with zero group velocity that we will discuss in section 6.5. According to the bifurcation diagram shown in Figure 6.5, we find contact defects before the bifurcation and sources afterward. Note that the source coexists with the small-amplitude sink that is created in the saddle-node bifurcation. Source and sink together can now form a bound state that corresponds to a transmission defect.

Other bifurcations. Defects may also arise when additional harmonic frequencies are introduced. Motivated by experiments [58] and numerical simulations [18], we have analyzed which defects can be created near period-doubling bifurcations of wave trains. We showed

[9] If $c_\mathrm{g}(k_\pm) < c_\mathrm{d}$, then we say, *by definition*, that the wave train at $\xi = -\infty$ is behind the defect, while the wave train at $\xi = \infty$ is ahead of the defect. If $c_\mathrm{g}(k_\pm) > c_\mathrm{d}$, then we reverse these definitions.

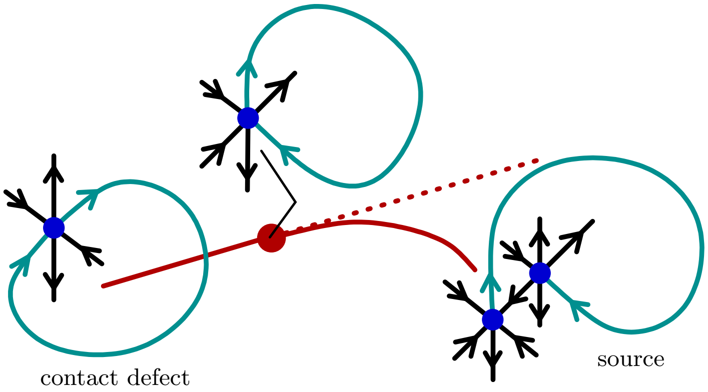


Figure 6.5. *The double-flip homoclinic bifurcation in the three-dimensional parameter space.*

in [52] that phase-slip defects can arise as interfaces between spatially homogeneous oscillations with a relative phase shift of π.

Of course, there are many more bifurcations that we expect to encounter. Sinks, for instance, persist even if we vary the two wavenumbers of the asymptotic wave trains independently, and we should therefore expect to observe any heteroclinic bifurcation of codimension two. In fact, even nontransverse homoclinic orbits may occur for large sets in parameter space.

The examples above notwithstanding, homoclinic and heteroclinic bifurcations can result in very complicated solution structures, and a complete classification appears to be impossible. To give a few examples, homoclinic orbits with complex spatial Floquet exponents are accompanied by a plethora of multiloop solutions. Sources and sinks that travel with the same speed form a heteroclinic loop, and the resulting heteroclinic bifurcation may lead to various different source-sink bound states (each being a transmission defect).

6.5. Pinning at inhomogeneities. Most of the counting arguments in sections 4 and 5 rely on the fact that $\partial_\tau u_\mathrm{d}$ and $\partial_\xi u_\mathrm{d}$ provide bounded solutions of the linearization of (4.2) about a defect, while ω_d and c_d provided the corresponding Lagrange multipliers. Spatial inhomogeneities break the translation symmetry, which prevents us from using moving frames. In particular, we will only be able to study defects with vanishing speed $c_\mathrm{d} = 0$. Thus, consider the equation

$$u_t = Du_{xx} + f(u) + \varepsilon g(x, u), \tag{6.8}$$

where we assume that the inhomogeneity g is localized so that $g(x, u) \to 0$ exponentially as $x \to \pm\infty$. The following result shows that standing sources persist at isolated positions for $\varepsilon \neq 0$ and are therefore pinned to the inhomogeneity. Similar results are true for contact defects with zero speed.

Theorem 6.2. *Suppose that $u_\mathrm{d}(x, \tau)$ is a transverse source of (1.1) with $c_\mathrm{d} = 0$. If we define*[10]

$$M(p) := \int_{-\infty}^{\infty}\int_0^{2\pi} \langle \psi_\mathrm{d}^c(x, \tau), g(x - p, u_\mathrm{d}(x, \tau))\rangle_{\mathbb{R}^n} \,\mathrm{d}\tau\,\mathrm{d}x,$$

[10] The functions $\psi_\mathrm{d}^c(x, \tau)$ and $\psi_\mathrm{d}^\omega(x, \tau)$ have been defined in Corollary 4.6.

then the source $u_{\mathrm{d}}(x+p,\tau)$ persists as a solution to (6.8) for ε close to zero provided p is a simple root of the Melnikov function $M(p)$. Furthermore, the temporal frequency $\omega_{\mathrm{d}}^*(\varepsilon)$ of the perturbed source is given by

$$\omega_{\mathrm{d}}^*(\varepsilon) = \omega_{\mathrm{d}} + \varepsilon \int_{-\infty}^{\infty}\int_0^{2\pi} \langle \psi_{\mathrm{d}}^{\omega}(x,\tau), g(x-p, u_{\mathrm{d}}(x,\tau))\rangle_{\mathbb{R}^n}\, \mathrm{d}\tau\, \mathrm{d}x + \mathrm{O}(\varepsilon^2).$$

Proof. The result follows from Lyapunov–Schmidt reduction applied to the spatial-dynamical system (4.2). We omit the details and refer instead to [39, section 5] and [49, section 8], where analogous analyses have been carried out. ■

Corollary 6.3. *If the hypotheses of the preceding theorem are met, and both the source and the inhomogeneity are symmetric (i.e., invariant under $x \mapsto -x$), then the Melnikov function $M(p)$ is odd. The symmetric source located at $x=0$ persists for $\varepsilon \neq 0$ provided*

$$\int_{-\infty}^{\infty}\int_0^{2\pi} \langle \psi_{\mathrm{d}}^{\mathrm{c}}(x,\tau), g_x(x, u_{\mathrm{d}}(x,\tau))\rangle\, \mathrm{d}\tau\, \mathrm{d}x \neq 0.$$

Proof. The bounded solution of (2.17) that corresponds to $\psi_{\mathrm{d}}^{\mathrm{c}}$ is of the form

$$\boldsymbol{\psi}(\xi,\tau) = \begin{pmatrix} c_{\mathrm{d}}\psi(\xi,\tau) - \partial_\xi \psi(\xi,\tau) \\ D\psi(\xi,\tau) \end{pmatrix}.$$

It is a consequence of [57] and the discussion in section 6.3 that $\boldsymbol{\psi}(0,\cdot)$ lies in the fixed-point space $\mathrm{Fix}(\mathcal{R}_0)$ of the reverser $\mathcal{R}_0$. In particular, $\psi_{\mathrm{d}}^{\mathrm{c}}(x,\tau)$ is odd in x, and the corollary follows then from Theorem 6.2. ■

In summary, large-amplitude sources with zero speed will be pinned to inhomogeneities, and their temporal frequency will change according to Theorem 6.2. In general, we therefore expect to see several pinned sources that have different temporal frequencies which depend on the location of the defects. Thus, our analysis seems to corroborate the statements made in [23, section 6.2.1], where inhomogeneities are mentioned as a possible explanation for the occurrence of what appears to be a one-parameter family of sources in the experiments [1]. The experimental observations reported in [36] seem to indicate, however, that these sources drift and are therefore not pinned. This issue therefore warrants further investigation.

For sinks with $c_{\mathrm{d}} = 0$, we do not expect pinning. Indeed, the intersection of center-unstable and center-stable manifolds is transverse for sinks at $\varepsilon = 0$ and therefore persists as a family of transverse intersections for all small ε independently of where we place the sink.

An alternative heuristic way of investigating the interaction of small-amplitude sinks with even smaller localized slowly varying inhomogeneities is via the approximation by the Burgers equation that we discussed in section 1.3. For $\varepsilon = 0$, the sinks or shocks in the Burgers equation have an eigenvalue at zero which is induced by translation symmetry. The resulting normally hyperbolic invariant manifold that consists of all translates of the sink persists for all ε close to zero. The leading-order terms of the perturbed flow on the invariant manifold are obtained by projecting the term in the Burgers equation that represents the inhomogeneous term $g(x,u)$ in (6.8) onto the manifold using the adjoint eigenfunction associated with the translation eigenvalue. In conservation laws, the null space of the adjoint is spanned by the constant function, so the perturbed flow is given by the mass of the term that represents the

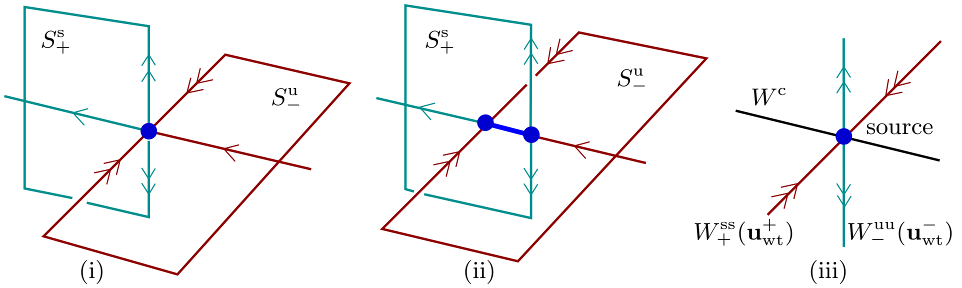


Figure 6.6. *Inhomogeneities may create contact defects* (ii) *or sources* (iii).

inhomogeneity $g(x,u)$. Since the Burgers equation (1.19) is written in terms of the wavenumber, it turns out that the derivative $g_x(x,u)$ of the inhomogeneity arises in the Burgers equation. Therefore, since the mass of $g_x(x,u)$ is zero, the reduced flow on the perturbed manifold vanishes to leading order, and the family of sinks persists with positions that are independent of the inhomogeneity. The spatial-dynamics argument presented above shows that the higher-order terms do not make a difference and that the situation remains unchanged for large-amplitude sinks and, typically, for large inhomogeneities.

This analysis of the Burgers equation suggests also that sinks with $c_{\mathrm{d}} \neq 0$ will simply outrun the inhomogeneity without experiencing a noticeable change of velocity so that the inhomogeneity will merely affect the asymptotic wave train. Thus, we shall now briefly discuss the influence of inhomogeneities on wave trains. Note that wave trains with nonzero group velocity are not affected by an inhomogeneity as they arise as transverse intersections of their center-stable and center-unstable manifolds of (6.8) with $\varepsilon = 0$ which persist for all ε close to zero. In this setting, we may interpret inhomogeneities as pinned transmission defects.

Wave trains with zero group velocity behave differently since small localized inhomogeneities create either contact defects or sources within such wave trains. To prove this, we again interpret wave trains as transverse intersections of center-unstable and center-stable manifolds. For wave trains with zero group velocity, this intersection, which consists precisely of the center manifold, is two-dimensional, and the vector field on it is given by (6.7) [11]. Factoring out the S^1-symmetry that is generated by the temporal time shift, we are left with a small line segment parametrized by the wavenumber q. As illustrated in Figure 6.6(i), this line segment, considered as a subset of W^{cu}_{-}, is separated at $\mathbf{u}^{\mathrm{uu}}_{-}$ into two half-lines by the strong-unstable manifold of the periodic orbit that corresponds to the wave train. Similarly, the strong-stable manifold cuts the line segment, considered as a subset of W^{cs}_{+}, into two half-lines at $\mathbf{u}^{\mathrm{ss}}_{+}$. Without an inhomogeneity, we have $\mathbf{u}^{\mathrm{ss}}_{+} = \mathbf{u}^{\mathrm{uu}}_{-} = \mathbf{u}_{\mathrm{wt}}$ as shown in Figure 6.6(i). The set of those initial data in W^{cu}_{-} whose solutions converge toward the wave train as $\xi \to -\infty$ (which we will refer to as the unstable set S^{u}_{-}) consists of points with line-segment coordinate $\mathbf{u} < \mathbf{u}^{\mathrm{uu}}_{-}$, i.e., of “half” of W^{cu}_{-}. Analogously, the stable set S^{s}_{+} of those data whose solutions converge to the wave train as $\xi \to \infty$ consists of points whose line-segment coordinate satisfies $\mathbf{u} > \mathbf{u}^{\mathrm{ss}}_{+}$.

Upon introducing the inhomogeneity, the transverse intersection will persist, but the stable and unstable sets may split. If $\mathbf{u}^{\mathrm{uu}}_{-} > \mathbf{u}^{\mathrm{ss}}_{+}$ as illustrated in Figure 6.6(ii), then an intersection

$$S^{\mathrm{s}}_{+} \cap S^{\mathrm{u}}_{-} = \{\mathbf{u} \in W^{\mathrm{cu}}_{-} \cap W^{\mathrm{cs}}_{+};\ \mathbf{u}^{\mathrm{uu}}_{-} > \mathbf{u} > \mathbf{u}^{\mathrm{ss}}_{+}\}$$

occurs which corresponds to a one-parameter family of contact defects with *varying* phase slip. If the stable and unstable sets split in the opposite direction, so that $\mathbf{u}_-^{\mathrm{uu}} < \mathbf{u}_+^{\mathrm{ss}}$, then contact defects cannot appear. In this case, we may, however, use the parameter $\omega = \omega_{\mathrm{nl}}(k_*) + \breve{\omega}$ to unfold the saddle-node wave trains at $x = \pm\infty$. While the intersection points on the fibers move linearly with $\breve{\omega}$, the dynamics on the base points of the fibers changes with the square root of $\breve{\omega}$. Thus, we denote the strong-stable manifold of the wave train $\mathbf{u}_{\mathrm{wt}}^+$ at $x = \infty$ with positive group velocity by $W_+^{\mathrm{ss}}(\mathbf{u}_{\mathrm{wt}}^+)$ and, analogously, the strong-unstable manifold of the wave train $\mathbf{u}_{\mathrm{wt}}^-$ at $x = -\infty$ with negative group velocity by $W_-^{\mathrm{uu}}(\mathbf{u}_{\mathrm{wt}}^-)$. Furthermore, we denote their intersections with the one-dimensional manifold W^{c} at $x = 0$ by $\mathbf{v}_+^{\mathrm{ss}}$ and $\mathbf{v}_-^{\mathrm{uu}}$, respectively. We then have $\mathbf{v}_+^{\mathrm{ss}}(\breve{\omega}) = \mathbf{u}^{\mathrm{ss}}(\varepsilon) - a_+\sqrt{\breve{\omega}}$ and $\mathbf{v}_-^{\mathrm{uu}}(\breve{\omega}) = \mathbf{u}^{\mathrm{uu}}(\varepsilon) + a_-\sqrt{\breve{\omega}}$ for certain positive coefficients $a_\pm > 0$. In particular, we find a unique $\breve{\omega}$ such that $\mathbf{u}_+^{\mathrm{ss}}(\breve{\omega}) = \mathbf{u}_-^{\mathrm{uu}}(\breve{\omega})$, which corresponds to a source as claimed.

6.6. Frequency locking through periodic forcing. Time-periodic forcing of the reaction-diffusion system breaks the temporal translation invariance and therefore tends to lock the frequency of defects to integer multiples of the forcing frequency ω_{f}. As a consequence, the frequency is effectively removed as a parameter, and the time-derivative $\partial_\tau u_{\mathrm{d}}$ of defects no longer contributes to the null space of Φ_{d}. Wave trains are still parametrized by their wavenumber k and arise as time-periodic solutions in a frame that moves with speed c, where

$$\omega_{\mathrm{nl}}(k) - kc = \omega_{\mathrm{f}}. \tag{6.9}$$

Note that the forcing frequency ω_{f} and the nonlinear dispersion relation ω_{nl} in the above equation can be replaced by rational multiples, which leads to further complications. For simplicity, we therefore focus on the simplest possible scenario. Thus, assume that the autonomous system (4.2) admits a transverse defect so that (6.9) is met for both $k = k_-$ and $k = k_+$ with $c = c_{\mathrm{d}}$. We then add a forcing term $\varepsilon g(t, u)$ with temporal frequency ω_{f} to (1.1). A sink will persist regardless of its phase relative to the periodic forcing since the map ι_{si} defined in (5.2) is onto (here, we add ε as a parameter and keep $k_\pm$ fixed). For sources, we cannot vary the variable ω_{d} that appears in the map ι_{so} from (5.4). Instead, we use the relative phase difference θ defined via $u_{\mathrm{d}}(\xi, \tau + \theta)$. As in section 6.5, we can then prove that sources persist for phase differences that correspond to simple roots of an appropriate Melnikov function $M(\theta)$. There should be an even number of such roots, which correspond to persisting sources that are alternately spectrally stable and unstable.

Alternatively, we may seek small-amplitude defects by studying the effect of temporal forcing on wave trains. As outlined above, wave trains will simply adjust their speed to ensure that (6.9) is met, and small-amplitude defects will therefore not occur. This works except when the forcing is in resonance with the group velocity so that

$$\omega_{\mathrm{nl}}(k_*) - k_* c_{\mathrm{g}}(k_*) = \omega_{\mathrm{f}}.$$

The spatial dynamics near the forced wave train is then described by an autonomous saddle-node bifurcation

$$\theta_\xi = k_* + \check{k} + \mathrm{O}(\varepsilon), \qquad \check{k}_\xi = \check{c}(k_* + \check{k}) - \check{k}^2 + \mathrm{O}(\varepsilon)$$

of a periodic orbit whose temporal frequency is locked to $\omega_{\rm f}$. The variables $\check{k}$ and $\check{c}$ denote the deviations from the wavenumber k_* and the group velocity $c_{\rm g}(k_*)$, respectively. Thus, we find small heteroclinic orbits that correspond to small-amplitude sinks. If we wish to find more complicated defects, we need to consider forcing frequencies that are in resonance with spatially homogeneous oscillations for which $k_* = 0$.

6.7. Locking of defect speed and phase velocity. Frequency locking can also occur when $\omega_{\rm d}$ vanishes at a given defect. In this case, the defect does not depend on time when considered in its comoving frame and therefore satisfies the travelling-wave ODE (2.4). We briefly discuss for each defect type with what codimension this situation arises. Note that $\omega_{\rm d} = 0$ if and only if the phase velocities $c_{\rm p}(k_-) = c_{\rm p}(k_+)$ are equal. Since we are mainly interested in waves with a nonzero group velocity, we assume that the nonlinear dispersion relation $\omega_{\rm nl}(k)$ is not constant on any open nonempty interval.

Transmission defects with $k_- = k_+$ occur as transverse homoclinic connections of a hyperbolic periodic orbit of (2.4). The periodic orbits as well as the homoclinic connection therefore persist if we vary $k_- = k_+$, provided we choose $c = c_{\rm p} = \omega_{\rm nl}(k)/k$. Thus, transmission defects arise with the same codimension as travelling waves and as proper modulated waves. In particular, defect speed and phase velocities can lock.

In contrast, phase velocity and defect speed of sinks, sources, and contact defects do not lock, since these defects arise as travelling waves only with a larger codimension than as defects with $\omega_{\rm d} \neq 0$. Indeed, sinks and sources have $k_- \neq k_+$, so the condition $c_{\rm p}(k_-) = c_{\rm p}(k_+)$ is a genuine additional equation that raises the codimension by one. For contact defects, we need $c_{\rm p}(k) = c_{\rm g}(k)$ along the branch which holds only if $\omega''_{\rm nl}(k) = 0$ as an explicit computation shows.

6.8. Large bounded domains. So far, we have focused on defects on the unbounded real line. Experiments, however, take place on bounded domains in a fixed laboratory frame. Thus, defects with nonzero speed are transient phenomena that disappear by colliding either with other defects or with the domain boundaries. Defects with speed close to zero, however, should exist over much longer time intervals. In this section, we shall therefore discuss the persistence of defects on large but bounded domains.

An additional motivation is provided by the need for computational tools that allow us to compute defects in an efficient way. Stable defects can, of course, be computed using direct simulations. If we are, however, interested in computing defects for many different parameter values, in finding their bifurcation points, or in computing unstable defects, then a formulation as a boundary-value problem that allows for a systematic continuation—using, for instance, AUTO97 [9]—would be desirable. To obtain such a formulation, we consider (4.2) in the frame of the defect, truncate the real line to a finite bounded interval, and impose appropriate phase and boundary conditions. We currently implement this procedure in AUTO97.

Boundary layers. We begin by investigating how wave trains interact with boundaries. Thus, consider the half line $\mathbb{R}^+$ with a boundary at $x = 0$. We focus on $c = 0$ and seek defects that are caused by the boundary conditions at $x = 0$. In the context of the modulated-wave equation (4.2), the boundary conditions at $x = 0$ are represented by the infinite-dimensional subspace B_- of Y that consists of all elements of Y which satisfy the boundary conditions.

We are interested in finding defects as orbits of (4.2) that lie in the intersection of B_- and the center-stable manifold $W^{\rm cs}_+(\mathbf{u}_{\rm wt})$ of a wave train $\mathbf{u}_{\rm wt}$. Thus, in analogy to section 6.3, we consider the map

$$\iota_{\rm bc}: \quad B_- \times W^{\rm cs}_+ \longrightarrow Y, \quad (\mathbf{u}^0, \mathbf{u}^+) \longmapsto \mathbf{u}^0 - \mathbf{u}^+.$$

For boundary conditions such as Dirichlet, Neumann, and various mixed conditions for which the reaction-diffusion system (1.1) is well-posed, the injection map $\iota_{\rm bc}$ is Fredholm, and its index is zero if $c_{\rm g} > 0$ and one if $c_{\rm g} \leq 0$.

Therefore, we expect one-parameter families of sinks for wave trains, with group velocity $c_{\rm g} < 0$ toward the boundary, that connect to the boundary at $x = 0$. This family of sinks is parametrized by the wavenumber k of the wave train at $x = \infty$. On the other hand, boundary-layer sources, which connect to wave trains with group velocity $c_{\rm g} > 0$ away from the boundary, occur only for isolated wavenumbers k. Similarly, there is typically only a finite number of boundary-layer contact defects since $c_{\rm g} = 0$ is fixed, even though the Fredholm index is one.

Examples of boundary-layer contact defects are homogeneous oscillations under Neumann boundary conditions. For small wavenumbers, and group velocities directed toward the boundary, we then find boundary-layer sinks which, for Neumann boundary conditions, can be thought of as symmetric sinks since we can extend the equation to the entire real line by reflecting across the boundary. Our discussion of inhomogeneities in section 6.5 can also be used to show that, for any homogeneous oscillation, there exists an open set, in fact a half space, of Robin boundary conditions that emit wave trains. The resulting defects are therefore boundary-layer sources. The boundary conditions in the complementary half space produce solutions that are asymptotically homogeneous oscillations, that is, boundary-layer contact defects. We also refer to the discussion in [55, section 5.3] on the existence of two-dimensional radially symmetric target patterns that are generated by boundary conditions imposed at a small hole in the domain.

Truncation. We are now prepared to discuss whether defects on $\mathbb{R}$ can be approximated by defects on large but bounded intervals $(-L, L)$. The discussion in the preceding section shows that the boundary conditions should be compatible in the sense that boundary-layer defects exist which absorb or generate the wave trains in the far field of the sources, contact defects, or sinks whose persistence we would like to prove.

Suppose, for instance, that there exist a source between wave trains with asymptotic wavenumbers $k_\pm$ and boundary-layer sinks that absorb these wave trains at the boundaries $x = \pm L$. As in [33], it then follows that the source persists as a solution to the truncated problem in a frame moving with speed $c(L)$ provided the frequency is adjusted to $\omega(L)$. Here, the corrections to speed and frequency satisfy $(c(L), \omega(L)) = (c_{\rm d}, \omega_{\rm d}) + {\rm O}({\rm e}^{-\delta L})$ for some $\delta > 0$. Furthermore, the solution is ${\rm O}({\rm e}^{-\delta L})$ close to the profile of the defect on $\mathbb{R}$ for $|\xi| < L/2$ and to the boundary-layer sinks for $|\xi| > L/2$. For symmetric sources and symmetric boundary conditions[11] at $\xi = \pm L$, we have $c(L) \equiv 0$.

[11] We say that the boundary conditions $\mathcal{B}_-$ and $\mathcal{B}_+$ at $\xi = -L$ and $\xi = L$ are symmetric if $\mathcal{R}_0(\mathcal{B}_-) = \mathcal{B}_+$ with $\mathcal{R}_0$ as in (6.6).

Similar results are true for sinks, although there is the severe restriction that both asymptotic wave trains need to be created by boundary-layer sources. For contact defects, we need to assume the existence of boundary-layer contact defects that connect to the boundary. To obtain persistence, we need to unfold the saddle-node wave trains in a fashion similar to the discussion in section 6.5. We omit the details.

Stability. Stability properties of defects on bounded intervals $(-L, L)$ with fixed boundary conditions are remarkably different from their counterparts on $\mathbb{R}$. First, the period map Φ_{d}^L of a defect on $(-L, L)$ will have only point spectrum. We denote the union of all Floquet exponents of Φ_{d}^L by Σ_L. If we take the limit as $L \to \infty$, then the set Σ_L converges to a limiting set in the symmetric Hausdorff distance where the convergence is uniform on bounded subsets of $\mathbb{C}$ [47]. The limiting set is the disjoint union of three sets, namely, the boundary spectrum Σ_{bdy}, the extended point spectrum Σ_{ext}, and the absolute spectrum Σ_{abs}. Before defining them in detail, we remark that Σ_{bdy} and Σ_{ext} are discrete, while Σ_{abs} is continuous in a sense that will be made precise below. The convergence of Σ_L toward both the boundary spectrum and the extended spectrum is exponential in L and includes convergence of the algebraic multiplicity. The convergence toward the absolute spectrum, however, is algebraic of order $\mathrm{O}(1/L)$, and the number of elements in Σ_L in any fixed neighborhood of an element of Σ_{abs} tends to infinity as $L \to \infty$.

We now define the three sets whose union is the limiting spectral set, and we begin with the absolute spectrum. We say that λ belongs to the complement of the absolute spectrum in $\mathbb{C}$ if there exist exponential weights $\eta_\pm$ such that $\Phi_{\mathrm{d}} - \rho$ is Fredholm with index zero on $L^2_{\eta_-,\eta_+}(\mathbb{R}, \mathbb{C}^n)$ *and* the relative Morse index of both asymptotic wave trains is zero relative to the exponential weights.[12] The absolute spectrum consists of a countable union of semi-algebraic curves which end precisely in spatial double roots ν with $\mathrm{Re}\, \nu = \eta_\pm$ of the dispersion relation of one of the asymptotic wave trains. We refer to [47, 46] for more details.

The extended point spectrum consists of all $\lambda \in \mathbb{C}$ for which there are exponential weights $\eta_\pm$ with the same properties as above so that the null space of $\Phi_{\mathrm{d}} - \rho$ on $L^2_{\eta_-,\eta_+}(\mathbb{R}, \mathbb{C}^n)$ is not trivial. Last, the boundary spectrum is defined as the extended point spectrum of the boundary-layer defects that are involved in the construction of the defect on $(-L, L)$.

Since we assumed that the wave trains are spectrally stable, we see that the absolute spectrum of sinks, transmission defects, and sources lies in the open left half-plane since the group velocities of the asymptotic wave trains are not zero. The absolute spectrum of contact defects always touches the imaginary axis at $\lambda = 0$ and consists, in fact, of a line segment $\lambda < 0$ inside a sufficiently small neighborhood of the origin. Indeed, the two roots $\nu_\pm$ near zero of the dispersion relation (6.2) $\lambda = d_\| \nu^2 + d_3 \nu^3 + \mathrm{O}(\nu^4)$ are complex conjugates for $\lambda < 0$.

The extended point spectrum of sinks, transmission defects, and sources contains $\lambda = 0$ with multiplicity

$$\begin{array}{lll} & 0 & \text{for sinks,} \\ (6.10) & 1 & \text{for transmission defects,} \\ & 2 & \text{for sources.} \end{array}$$

[12] This means that the relative Morse indices are computed for (4.6) with $\eta = \eta_\pm$ and $\mathbf{u}_{\mathrm{d}}$ replaced by the asymptotic wave train.

Last, the boundary spectrum near the origin is given as follows:

$$\begin{array}{ll} 0 & \text{for each boundary-layer sink,} \\ 1 & \text{for each boundary-layer source,} \end{array} \tag{6.11}$$

where the time-derivative of the boundary-layer source provides the eigenfunction.

Thus, in summary, if we consider the union of boundary and extended point spectrum near the origin for truncated sinks and sources, then both of them have two eigenvalues near the origin. For truncated sinks, these two eigenvalues arise due to the two boundary-layer sources near the boundaries $x = \pm L$, whereas the eigenvalues for truncated sources occur from the source itself with the two boundary-layer sinks not contributing any eigenvalues. While one of the two eigenvalues for truncated sinks or sources accounts for the time-derivative, i.e., for the temporal translation symmetry that is still present, the other eigenvalue, which is associated with the position of the defect relative to the boundary, is free to move and may stabilize or destabilize the truncated defect.

6.9. Interactions of defects. Heuristically, the interaction of defects with each other or with boundaries can be explained to a large extent by deriving formal solvability conditions. These conditions arise when we try to match defects, and they can be calculated by projecting certain matching terms onto the null space of $\Phi_{\rm d} - 1$, i.e., onto the eigenfunctions associated with $\lambda = 0$, using the adjoint eigenfunctions [14, 13, 43].

First, we need to clarify what we mean when we refer to the eigenfunctions of $\Phi_{\rm d}$ associated with the Floquet exponent $\lambda = 0$, since the essential spectrum of $\Phi_{\rm d}$ touches $\lambda = 0$. We believe that the eigenfunctions and eigenvalues that we need to take into account are those that arise in the weighted spaces given in Theorem 6.1. Indeed, the essential spectrum of defects is generated by the asymptotic wave trains, and its effect is accounted for by the Burgers equation (1.19). Given our spatial-dynamics interpretation of the group velocities, we believe that the interaction of defects manifests itself in their spectra in the spaces $L^2_\eta(\mathbb{R}, \mathbb{R}^n)$ with η chosen as in Theorem 6.1.

Transverse sinks do not have any eigenvalues at $\lambda = 0$ (see Theorem 6.1). Therefore, they do not interact with the tails of adjacent sources but instead react passively by adjusting their speed via (1.8) according to changes of the wavenumbers in the far field. Changing the phase (6.3) of one of the asymptotic wave trains will cause the sink to correct its position and temporal phase according to the algebraic phase jump condition (6.4).

Transverse sources have a double eigenvalue at $\lambda = 0$ that is induced by the derivatives of the source with respect to time and space. Therefore, for each source, we obtain two differential equations that define the time derivatives of its position and its phase as functions of perturbations in the far field. Since the associated adjoint eigenfunctions at $\lambda = 0$ are exponentially localized by Theorem 6.1, sources interact with adjacent defects or boundaries only very weakly, namely through terms of the form $\mathrm{e}^{-\delta L}$ for $\delta > 0$ that are exponentially small in the distance L between the source and other defects or boundaries.

Transmission defects have a simple eigenvalue at $\lambda = 0$, and the associated adjoint eigenfunction is localized behind the defect, where the group velocity points away from the interface, and constant ahead of the defect. We therefore believe that transmission defects will adjust their phase instantaneously whenever the phase of the wave train ahead of them is changed.

The position and temporal phase of transmission defects are also affected by the presence of boundaries or perturbations in their wake. These interactions are described by a single differential equation. Note that position and temporal phase are implicitly related through phase matching with the wave trains ahead of the transmission defect (see sections 6.1 and 6.2).

To our knowledge, interactions that involve contact defects have not been studied previously (not even on a formal level). We believe that Theorem 4.5, which asserts that the Evans function has a simple root at $\lambda = 0$, may play an important role, since this root may change the temporal algebraic decay rate of localized perturbations [35].

6.10. Genericity. Our goal in this paper has been to present a list of transverse defects. It is a challenging task to prove whether, and in what sense, this list is complete.

From a formal point of view, we included all possible combinations of group velocities $c_\mathrm{g}^\pm$ relative to the defect velocity c_d with the exception of *one-sided contact defects* for which one of the group velocities is equal to the defect speed but the other one is not, i.e., for which either $c_\mathrm{g}^- = c_\mathrm{d} \neq c_\mathrm{g}^+$ or $c_\mathrm{g}^- \neq c_\mathrm{d} = c_\mathrm{g}^+$. The reason for omitting one-sided contact defects is that they form part of the boundary of the region in (k_-, k_+) space where transverse sinks exist. Put differently, in any given system, we expect that only a finite number of wavenumbers occur that are generated by boundaries or by sources. In general, there is no reason why one-sided defects should be selected as they occur only in one-parameter families. For the same reasons, we do not take degenerate sinks, transmission defects, or contact defects into account even though degeneracies, such as eigenvalues at the origin with higher multiplicity, can typically be found by varying only the asymptotic wavenumbers. If we therefore wish to exclude degenerate defects as well as one-sided contact defects, we may consider introducing a notion of genericity that requires the persistence of generic defects when we truncate the real line to a large but finite interval with *generic* boundary layers in the sense of section 6.8.

The reason for including sources as generic defects is that they actively select wavenumbers. Similarly, we wish to regard contact and transmission defects as generic since they occur within a background of wave trains with identical wavenumber. In other words, the reason for their generic existence is that they accommodate phase slips within an oscillatory medium (see section 6.2). Interpreted in a different way, contact and transmission defects occur in open sets of wavenumbers once we impose the constraint that $k_- = k_+$.

From a purely mathematical viewpoint, and thinking solely in terms of the codimension with which a certain heteroclinic orbit exists, there is no difference between, say, transmission defects and one-sided contact defects. However, the goal would be to find a notion of genericity that accurately reflects the physical intuition that we described above, i.e., that allows sources, for instance, but rejects one-sided contact defects.

6.11. Higher space dimensions. The classification we have presented is valid for *essentially* one-dimensional media. In particular, our approach, and therefore our results, apply equally well to problems on cylindrical domains with a bounded higher-dimensional cross section. Indeed, the key feature that we exploited are exponential dichotomies which exist provided the operators encountered enjoy certain compactness properties that are satisfied for the cross-sectional variables whenever the cross section is bounded [39, 49].

In fact, our results can also be adapted immediately to cover line defects in the plane that are spatially periodic in the direction parallel to the line defect [52]. If we, for the sake of clarity, choose coordinates so that the line defect corresponds to $x = 0$, so that y denotes the variable parallel to the line defect, then the modulated-wave equation (4.2) would contain the Laplace operator in the y variable, while the spatial evolution variable would be the x-variable transverse to the defect. We would then pose (4.2) on the space of functions that are periodic in time and in y. For instance, line defects between planar wave trains that propagate in a direction parallel to the defect can be easily analyzed in this framework.

Certain two-dimensional point defects, such as spiral waves and target patterns in the Belousov–Zhabotinsky reaction or stationary focus patterns in convection experiments, are amenable to a similar description since we can measure group velocities in the radial variable [53, 55]. A classification analogous to the one presented here has not yet been attempted in higher space dimensions.

7. Example: The cubic-quintic Ginzburg–Landau equation. In this section, we illustrate that all four elementary transverse defects arise in the complex cubic-quintic Ginzburg–Landau equation (CQGL) for appropriate values of the coefficients. In fact, the existence of sources, sinks, and transmission defects in the CQGL has been established in [10], and we therefore focus here on the existence of contact defects. Thus, consider the CQGL

$$\check{A}_t = (1+\mathrm{i}\check{\alpha})\check{A}_{xx} + \check{A} - (1+\mathrm{i}\check{\gamma})\check{A}|\check{A}|^2 - (\check{\delta}_0+\mathrm{i}\check{\delta})\check{A}|\check{A}|^4, \tag{7.1}$$

where $x \in \mathbb{R}$, $\check{A}(x,t) \in \mathbb{C}$, and all coefficients are real. The CQGL arises as a special case of the modulation equation that describes degenerate Hopf bifurcations in reaction-diffusion systems [12] and, for $\check{\alpha} = 0$, coincides with the λ-ω system that has been investigated, for instance, in [32]. The CQGL respects the gauge symmetry $\check{A} \mapsto \mathrm{e}^{\mathrm{i}\phi}\check{A}$ so that we should seek relative equilibria of the form

$$\check{A}(x,t) = A(x-\check{c}t)\mathrm{e}^{-\mathrm{i}\check{\omega}t}.$$

Substituting this ansatz into (7.1) yields the ODE

$$A'' = -\frac{1}{1+\mathrm{i}\check{\alpha}}\left[(1+\mathrm{i}\check{\omega})A + \check{c}A' - (1+\mathrm{i}\check{\gamma})A|A|^2 - (\check{\delta}_0+\mathrm{i}\check{\delta})A|A|^4\right]. \tag{7.2}$$

We choose $1+\check{\alpha}\check{\gamma} > 0$, which allows us to rescale A and the parameters so that (7.2) becomes

$$A'' = -(1+\mathrm{i}\omega)A - cA' + (1+\mathrm{i}\gamma)A|A|^2 + (\delta_0+\mathrm{i}\delta)A|A|^4. \tag{7.3}$$

For the sake of clarity, we set $\delta_0 = 0$ from now on but remark that most of the subsequent analysis remains valid, with appropriate modifications, when $\delta_0 > 0$. An essential assumption is that we take the remaining parameters $(\gamma, \delta, \omega, c)$ to be close to zero, so that we perturb from the real CGL.

In summary, we consider the complex two-dimensional ODE

$$\begin{aligned} A' &= B, \\ B' &= -(1+\mathrm{i}\omega)A - cB + (1+\mathrm{i}\gamma)A|A|^2 + \mathrm{i}\delta A|A|^4 \end{aligned} \tag{7.4}$$

with small external parameters (γ, δ) and small internal parameters (ω, c). In passing, we remark that (7.4) has the same structure as the modulated-wave equation (4.2). In fact, regardless of whether (7.4) is derived near a degenerate Hopf bifurcation or represents a λ-ω system, (7.4) *is* the modulated-wave equation, restricted to either a center manifold or a finite-dimensional Fourier subspace. In each case, the gauge invariance is induced by the time-shift symmetry and is therefore a genuine symmetry rather than a normal-form symmetry. For $c = 0$, (7.4) has two reversers given by

$$\mathcal{R}_0 : (A, B) \longmapsto (A, -B), \qquad \mathcal{R}_\pi : (A, B) \longmapsto (-A, B).$$

As already alluded to, we study (7.4) as a perturbation from the real Ginzburg–Landau equation

$$\begin{aligned} A' &= B, \\ B' &= -A + A|A|^2. \end{aligned} \tag{7.5}$$

Note that (7.5) is a Hamiltonian system where the Hamiltonian H and the symplectic matrix J are given by

$$H = |A|^2 + |B|^2 - \frac{1}{2}|A|^4, \qquad J(A, B, \bar{A}, \bar{B}) = (\bar{B}, -\bar{A}, B, -A).$$

The real Ginzburg–Landau equation (7.5) has a family

$$A(x; k) = \sqrt{1 - k^2}\mathrm{e}^{\mathrm{i}kx} \tag{7.6}$$

of wave trains for $|k| < 1$. In particular, the dispersion relation $\omega = \omega_{\mathrm{nl}}(k)$ is degenerate as all these wave trains have $\omega = 0$. There also is an explicit defect given by

$$A(x) = \tanh\left(\frac{x}{\sqrt{2}}\right), \tag{7.7}$$

which connects the wave trains with zero wavenumber with a phase slip of π (see section 6.2).

Once γ or δ are nonzero, the wave trains (7.6) are solutions to (7.4) if and only if

$$\omega = \omega_{\mathrm{nl}}(k; \gamma, \delta, c) = (1 - k^2)\gamma + (1 - k^2)^2\delta - kc, \tag{7.8}$$

which is the dispersion relation in the frame moving with speed c. The wave trains disappear in saddle-node bifurcations precisely when

$$c = c_{\mathrm{g}}(k; \gamma, \delta) = -2k\gamma + 4k(k^2 - 1)\delta. \tag{7.9}$$

We are interested in the fate of the defect (7.7) for (γ, δ) small but nonzero.

It will be convenient to eliminate the gauge symmetry of (7.5). Often, this is achieved by introducing the blow-up coordinates B/A (or A/B) and $|A|^2$. In our example, however, this would force us to use two blow-up charts since both A and B vanish somewhere along the defect. Instead, we use the generators of the ring of invariant polynomials with respect to the S^1-gauge symmetry as new coordinates and define

$$R = A\bar{A}, \qquad S = B\bar{B}, \qquad N = A\bar{B}, \tag{7.10}$$

with $N = N_\mathrm{r} + \mathrm{i}N_\mathrm{i}$. A result by Hilbert (see, for instance, [5]) shows that the invariants smoothly parametrize the group orbits of the S^1-symmetry. For later use, we remark that, if we evaluate N on the wave trains (7.6), we obtain $N = A\bar{A}' = -\mathrm{i}k(1-k^2)$ so that

$$N_\mathrm{i} = -k(1-k^2). \tag{7.11}$$

In the new variables (7.10), equation (7.5) becomes

$$\begin{aligned} R' &= 2N_\mathrm{r}, \\ S' &= 2(R-1)N_\mathrm{r} - 2cS + 2(\omega - \gamma R - \delta R^2)N_\mathrm{i}, \\ N_\mathrm{r}' &= S - cN_\mathrm{r} - R + R^2, \\ N_\mathrm{i}' &= -cN_\mathrm{i} + \omega R - \gamma R^2 - \delta R^3. \end{aligned} \tag{7.12}$$

Note that we eliminated the phase invariance without reducing the dimension, at the expense of introducing an algebraic relation

$$\mathcal{C}(R,S,N) := RS - N\bar{N} = 0, \tag{7.13}$$

which must be satisfied for solutions of (7.12) to correspond to solutions of (7.5). For $c = 0$, (7.12) is reversible under $N \mapsto -N$ (which represents both $\mathcal{R}_0$ and $\mathcal{R}_\pi$).

For $(\gamma, \delta) = (\omega, c) = 0$, we find

$$\begin{aligned} R' &= 2N_\mathrm{r}, \\ S' &= 2(R-1)N_\mathrm{r}, \\ N_\mathrm{r}' &= S - R + R^2, \\ N_\mathrm{i}' &= 0. \end{aligned} \tag{7.14}$$

We note that both $\mathcal{C}$ and H are conserved along trajectories of (7.14). Exploiting that the Hamiltonian is now given by

$$H = S + R - \frac{R^2}{2}, \tag{7.15}$$

(7.14) can therefore be written as

$$R'' = 2H - 4R + 3R^2, \qquad S = H - R + R^2/2, \qquad N = R'/2. \tag{7.16}$$

The equilibria of (7.14) are given by $S = R - R^2$ and $N_\mathrm{r} = 0$ for arbitrary R and N_i. Using (7.11) and the algebraic relation (7.13), we obtain the parametrization

$$(R^\infty, S^\infty, N_\mathrm{r}^\infty, N_\mathrm{i}^\infty) = (1-k^2)(1, k^2, 0, -k), \qquad H^\infty = \frac{(1-k^2)(1+3k^2)}{2} \tag{7.17}$$

for those equilibria of (7.12) that correspond to solutions of (7.5). The eigenvalues of the linearization of (7.14) about the equilibria (7.17) are given by two zero eigenvalues, $\lambda_1 = \lambda_2 = 0$, which arise due to translation symmetry and the conserved quantity $\mathcal{C}$, and two eigenvalues $\lambda_{3/4} = \pm\sqrt{2(1-3k^2)}$, which are hyperbolic if $k^2 < 1/3$. Since the waves with $k^2 > 1/3$ are known to be unstable with respect to long-wavelength perturbations, the well-known Eckhaus instability, we focus exclusively on the case

$$|k| < \frac{1}{\sqrt{3}} \tag{7.18}$$

so that we are away from the Eckhaus instability.

On the other hand, solving (7.16), we find the one-parameter family

$$\begin{aligned} R_{\mathrm{d}}(x) &= (1-k^2) - (1-3k^2)\,\mathrm{sech}^2\left(\sqrt{1-3k^2}x/\sqrt{2}\right), \\ S_{\mathrm{d}}(x) &= H^\infty - R_{\mathrm{d}}(x) + \frac{1}{2}R_{\mathrm{d}}^2(x), \\ N_{\mathrm{d,r}}(x) &= R_{\mathrm{d}}'(x)/2, \\ N_{\mathrm{d,i}}(x) &= -k(1-k^2) \end{aligned} \tag{7.19}$$

of homoclinic orbits that are parametrized by the asymptotic wavenumber k. Linearizing (7.14) about these orbits gives the variational equation

$$\begin{aligned} R' &= 2N_{\mathrm{r}}, \\ S' &= 2(R_{\mathrm{d}}(x)-1)N_{\mathrm{r}} + R_{\mathrm{d}}'(x)R, \\ N_{\mathrm{r}}' &= S - R + 2R_{\mathrm{d}}(x)R, \\ N_{\mathrm{i}}' &= 0 \end{aligned} \tag{7.20}$$

and the corresponding adjoint variational equation

$$\begin{aligned} r' &= -R_{\mathrm{d}}'(x)s - (2R_{\mathrm{d}}(x)-1)n_{\mathrm{r}}, \\ s' &= -n_{\mathrm{r}}, \\ n_{\mathrm{r}}' &= -2r - 2(R_{\mathrm{d}}(x)-1)s, \\ n_{\mathrm{i}}' &= 0. \end{aligned} \tag{7.21}$$

Solutions to the adjoint variational equation (7.21) can be computed explicitly in various different ways. One approach is to observe and exploit that

$$[s']'' = (6R_{\mathrm{d}}(x)-4)[s'].$$

Alternatively, the gradients $\nabla\mathcal{C}$ and ∇H of the conserved quantity (7.13) and the energy (7.15) are automatically solutions to (7.21). Either way, we see that three bounded solutions to (7.15) are given by

$$\begin{aligned} \psi_0(x) &= \left(\frac{1}{2}R_{\mathrm{d}}^2(x) + (R^\infty-1)R_{\mathrm{d}}(x) + R^\infty - \frac{3}{2}(R^\infty)^2, R_{\mathrm{d}}(x) - R^\infty, -R_{\mathrm{d}}'(x), 0\right), \\ \psi_1(x) &= (0,0,0,1), \\ \psi_2(x) &= \nabla H(x) \;=\; (1-R_{\mathrm{d}}(x),1,0,0)\,, \end{aligned}$$

while the fourth, linearly independent solution $\psi_3(x)$ is unbounded. Note that $\psi_0(x)$ decays exponentially and points to the direction perpendicular to the sum of the center-stable

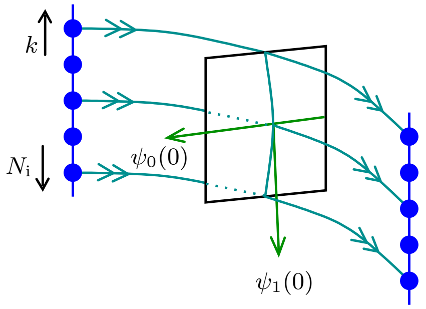


Figure 7.1. *The dynamics of* (7.14) *within the level set* $\mathcal{C} = 0$ *is illustrated. The two lines of equilibria are drawn separately even though they both correspond to the same family given in* (7.22).

and center-unstable manifolds along the homoclinic orbit (7.19). The two vectors $\psi_0(x)$ and $\psi_1(x)$ together are part of the orthogonal complement of the strong-unstable manifold, which coincides with the strong-stable manifold. We refer to Figure 7.1 for an illustration.

Our goal is to understand the intersections of various stable and unstable manifolds as well as their dependence on the wavenumber k of the wave trains. Their intersections will be studied using the Melnikov integrals in the directions of ψ_0 and ψ_1 for the derivatives of the right-hand side of (7.12) with respect to $(\gamma, \delta, \omega, c)$. The resulting bifurcation scenario is, in fact, similar to the double-flip bifurcation that we discussed in sections 6.4 and 6.5 in the context of transitions from contact defects to sources and of inhomogeneities embedded in media of wave trains with zero group velocity.

We begin with a local analysis of the slow manifold

$$(R^\infty, S^\infty, N_{\mathrm{r}}^\infty, N_{\mathrm{i}}^\infty) = \big(1 - k^2, k^2(1 - k^2), 0, -k(1 - k^2)\big) \tag{7.22}$$

that consists of all equilibria of (7.14) that satisfy (7.13). The tangent space to this family of equilibria is spanned by

$$\frac{\mathrm{d}}{\mathrm{d}k}(R^\infty, S^\infty, N_{\mathrm{r}}^\infty, N_{\mathrm{i}}^\infty) = \big(-2k, 2k(1 - 2k^2), 0, 3k^2 - 1\big)\,, \tag{7.23}$$

while the associated adjoint eigenvector is

$$\psi_1^\infty = (0, 0, 0, 1).$$

The flow along the family of equilibria for $(\gamma, \delta, \omega, c) \neq 0$ is obtained [16] by evaluating the derivatives of the right-hand side of (7.12) with respect to $(\gamma, \delta, \omega, c)$ at the equilibria and projecting them with ψ_1^∞. If we parametrize the center manifold by the wavenumber k, the reduced equation is

$$(3k^2 - 1)\dot{k} = ck(1 - k^2) + \omega(1 - k^2) - \gamma(1 - k^2)^2 - \delta(1 - k^2)^3 + \mathrm{O}(2), \tag{7.24}$$

where $\mathrm{O}(2)$ denotes quadratic terms in $(\gamma, \delta, \omega, c)$. A point k_* is an equilibrium if and only if ω is given by the nonlinear dispersion relation (7.8)

$$\omega = \omega_{\mathrm{nl}}(k_*; \gamma, \delta, c) = \gamma(1 - k_*^2) + \delta(1 - k_*^2)^2 - ck_*, \tag{7.25}$$

in which case (7.24) becomes

$$\dot{k} = -\frac{1-k^2}{1-3k^2}[\omega_{\mathrm{nl}}(k_*;\gamma,\delta,c) - \omega_{\mathrm{nl}}(k;\gamma,\delta,c)] + \mathrm{O}(2).$$

The linearization of (7.24) about an equilibrium k_* is therefore not hyperbolic precisely when c is given by the group velocity (7.9)

$$\begin{aligned} c &= c_{\mathrm{g}}(k_*;\gamma,\delta) = -2k_*\gamma - 4k_*(1-k_*^2)\delta, \\ \omega &= \omega_{\mathrm{nl}}(k_*;\gamma,\delta,c_{\mathrm{g}}(k_*)) = (1+k_*^2)\gamma + (1-k_*^2)(1+3k_*^2)\delta. \end{aligned} \tag{7.26}$$

With these choices of ω and c, the slow reduced equation near $k = k_*$ is given by

$$\dot{\kappa} = -\frac{1-k_*^2}{1-3k_*^2}\left[\gamma + 2\delta(1-3k_*^2)\right]\kappa^2 + \mathrm{O}(\kappa^3) \tag{7.27}$$

in the variable $\kappa = k - k_*$. In particular, to leading order in κ, the unstable manifold W^{u} of the equilibrium $k = k_*$ is contained in $k < k_*$ if $\gamma + 2\delta > 0$ and in $k > k_*$ if $\gamma + 2\delta < 0$.

We now compute the Melnikov integrals in the direction of $\psi_0(x)$ which are defined by

$$M_0^j = \int_{\mathbb{R}} \langle \psi_0(x), \partial_j F(R_{\mathrm{d}}, S_{\mathrm{d}}, N_{\mathrm{d,r}}, N_{\mathrm{d,i}})(x)\rangle \,\mathrm{d}x,$$

where F denotes the right-hand side of (7.12) and $j = \gamma,\delta,\omega,c$. These integrals show how the center-stable and center-unstable manifolds split along the homoclinic orbit given in (7.19) upon varying (γ,δ,ω,c) near zero. A tedious calculation gives

$$\begin{aligned} M_0^\gamma(k) &= -\frac{4}{3}k(1-k^2)(1+3k^2)\sqrt{2(1-3k^2)}, \\ M_0^\delta(k) &= -\frac{4}{15}k(1-k^2)(3+2k^2+27k^4)\sqrt{2(1-3k^2)}, \\ M_0^\omega(k) &= 4k(1-k^2)\sqrt{2(1-3k^2)}, \\ M_0^c(k) &= \frac{4}{3}(1-k^2)\sqrt{2(1-3k^2)}. \end{aligned} \tag{7.28}$$

The distance of the center-unstable and center-stable manifolds is given by the expression

$$\Delta_0(k;\gamma,\delta,\omega,c) = \omega M_0^\omega(k) + cM_0^c(k) + \gamma M_0^\gamma(k) + \delta M_0^\delta(k) + \mathrm{O}(2).$$

If we seek contact defects, we should find intersections related to a saddle-node equilibrium k_* of (7.26). Thus, we shall choose (ω,c) according to (7.26) and investigate roots of the splitting distance

$$\begin{aligned} \Delta_0^{\mathrm{cd}}(k,k_*;\gamma,\delta) &= \omega_{\mathrm{nl}}(k_*)M_0^\omega(k) + c_{\mathrm{g}}(k_*)M_0^c(k) + \gamma M_0^\gamma(k) + \delta M_0^\delta(k) + \mathrm{O}(2) \\ &= \frac{4}{3}(1-k^2)(2-3kk_*-3k^2)(k-k_*)\sqrt{2(1-3k^2)}\,\gamma \\ &\quad + \frac{4}{15}(1-k^2)\left(k[12-2k^2-27k^4] - k_*\left[20(1-k_*^2) + 15kk_*(3k_*^2-2)\right]\right) \\ &\quad \times \sqrt{2(1-3k^2)}\,\delta + \mathrm{O}(2) \end{aligned} \tag{7.29}$$

of the center-unstable and center-stable manifolds of the equilibrium k_* along the homoclinic orbit at level k.

Next, we investigate the splitting of the strong-stable and strong-unstable manifolds in the center direction which is given by Melnikov integrals in the direction of $\psi_1(x)$. We therefore choose ω and c as in (7.26) so that a given point $k = k_*$ on the manifold of equilibria persists as a saddle-node equilibrium for $(\gamma, \delta) \neq 0$. We define the Melnikov integrals

$$\begin{aligned} M_1^\gamma(k_*) &= \int_{\mathbb{R}} \big\langle \psi_1(x), [\partial_\gamma + (1+k_*^2)\partial_\omega - 2k_*\partial_c] F(R_{\mathrm{d}}, S_{\mathrm{d}}, N_{\mathrm{d,r}}, N_{\mathrm{d,i}})(x) \big\rangle \,\mathrm{d}x, \\ M_1^\delta(k_*) &= \int_{\mathbb{R}} \big\langle \psi_1(x), [\partial_\delta + (1-k_*^2)(1+3k_*^2)\partial_\omega - 4k_*(1-k_*^2)\partial_c] \\ &\qquad\qquad \times F(R_{\mathrm{d}}, S_{\mathrm{d}}, N_{\mathrm{d,r}}, N_{\mathrm{d,i}})(x) \big\rangle \,\mathrm{d}x, \end{aligned}$$

where F denotes again the right-hand side of (7.12). A tedious but straightforward computation gives

$$M_1^\gamma(k_*) = \frac{2}{3}(1-3k_*^2)\sqrt{2(1-3k_*^2)} > 0, \qquad M_1^\delta(k_*) = \frac{16}{15}(1-3k_*^2)^2\sqrt{2(1-3k_*^2)} > 0. \tag{7.30}$$

We are now ready to describe the bifurcation picture for small (γ, δ). Consider a two-dimensional section transverse to the flow that lies in a small neighborhood of $(R_{\mathrm{d}}, S_{\mathrm{d}}, N_{\mathrm{d,r}}, N_{\mathrm{d,i}})(0)$ in the three-dimensional manifold described by the algebraic relation (7.13). In this two-dimensional section, the center-stable and center-unstable manifolds coincide when $(\gamma, \delta, \omega, c) = 0$. Their intersection forms a line which is parametrized by the coordinate k of the base point of the corresponding strong-stable fiber on the center manifold (which is, in fact, the manifold of equilibria). Since the general bifurcation diagram is rather complicated, we focus separately on the two cases $\delta = 0$ (the CGL) and $k_* = 0$.

First, we set $\delta = 0$. Since we are interested in contact defects, we choose (ω, c) according to (7.26) so that the fixed base point k_* persists as a saddle-node equilibrium. To find contact defects, we need to solve $\Delta_0^{\mathrm{cd}}(k, k_*; \gamma, 0) = 0$. Using its definition (7.29), we see that the nontrivial roots of this equation are given by $k = \pm 1/\sqrt{3}$ and by $k = k_*$. The first case is close to the Eckhaus instability, and we will not consider it here. Instead, we show that the second case cannot lead to contact defects: We focus first on the region $\gamma > 0$, so that the unstable manifold of the saddle-node equilibrium k_* for (7.27) is the set $k < k_*$. Since $M_1^\gamma(k_*) > 0$, however, the strong-unstable manifold of the equilibrium k_* will miss the stable set of the equilibrium k_* as illustrated in Figure 7.2(i). For $\gamma < 0$, both the sign of the splitting distance and the flow on the slow manifold change sign, and we again conclude that there is no homoclinic orbit to $k_* = 0$. In particular, there are no contact defects for $\delta = 0$. Varying ω and c to unfold the saddle node at $k = k_*$, we see that sources are created. Indeed, we can parametrize the equilibria in the unfolding of the saddle node using $\mu_1 := \sqrt{|\omega - \omega_{\mathrm{nl}}(k_*)|}$ or $\mu_2 := \sqrt{|c - c_{\mathrm{g}}(k_*)|}$. The Melnikov integrals with respect to μ_j vanish since the parameters μ_j enter only at second order. The base points, however, change to leading order in μ_j, which allows us to find an intersection of the strong-stable and strong-unstable fibers. These intersections, which correspond to sources, are known as the Nozaki–Bekki holes and can be given explicitly in terms of elementary functions [2, 41].

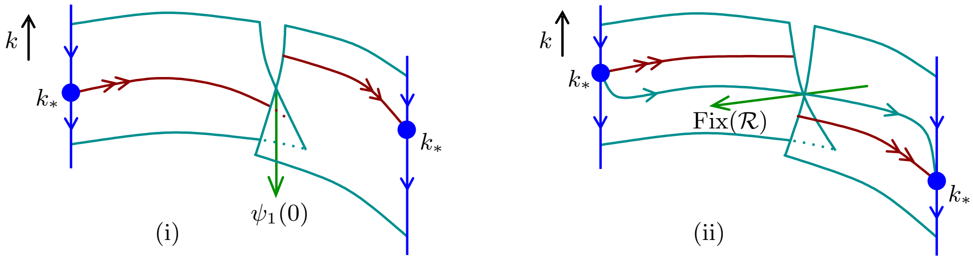


Figure 7.2. *The splitting of the strong-unstable and strong-stable manifolds of the equilibrium* $k_* = 0$ *of* (7.12) *is illustrated for* $\gamma + \frac{8}{5}\delta > 0$ *in* (i) *and for* $\gamma + \frac{8}{5}\delta < 0$ *in* (ii)*. The sketch of the flow on the slow manifold assumes* $\gamma + 2\delta > 0$*.*

Next, we set $k_* = 0$. In particular, $c = c_{\rm g} = 0$ by (7.26), which makes (7.12) reversible under $\mathcal{R} : N \mapsto -N$ with fixed-point space $\{N = 0\}$. The unstable manifold of the equilibrium $k_* = 0$ of (7.27) is the set $k < 0$ when $\gamma + 2\delta > 0$ and the set $k > 0$ for $\gamma + 2\delta < 0$. The splitting distance of the strong-unstable and the strong-stable manifold, on the other hand, is given by

$$\begin{aligned}\Delta_1^{\rm cd}(\gamma,\delta) &= \langle \psi_1(0), W^{\rm uu}(0) - W^{\rm ss}(0)\rangle = M_1^{\gamma}(0)\gamma + M_1^{\delta}(0)\delta + {\rm O}(2)\\ &\overset{(7.30)}{=} \frac{2\sqrt{2}}{3}\left[\gamma + \frac{8}{5}\delta\right] + {\rm O}(2).\end{aligned}$$

As shown in (7.23) and indicated in Figure 7.1, the vector $\psi_1(0)$ points to the positive $N_{\rm i}$ and the negative k direction. Thus, the strong-unstable and stable manifolds of $k_* = 0$ split as shown in Figure 7.2. We therefore need

$$(\gamma + 2\delta)\left(\gamma + \frac{8}{5}\delta\right) + {\rm O}(3) < 0 \tag{7.31}$$

to get a reversible homoclinic connection of the saddle-node equilibrium $k_* = 0$ as in Figure 7.2(ii). The inequality (7.31) defines, to leading order, a small nonempty cone in the (γ,δ)-plane (see Figure 7.3). The reversible connection indeed exists as both the center-unstable and the center-stable manifolds intersect the fixed-point space $\{N = 0\}$ of the reverser transversely, which follows from the expression for $\psi_1(0)$ and since $N_{\rm d}'(0) = R_{\rm d}''(0)/2 \neq 0$. Last, we show that the reversible contact defects persist if we vary the wavenumber k_* near $k_* = 0$. To accomplish this, it suffices to prove that the center-unstable and center-stable manifolds of the equilibrium $k_* = 0$ intersect transversely along the contact defect. Thus, we evaluate their splitting distance (7.29) at $k_* = 0$, which yields

$$\begin{aligned}\Delta_0^{\rm cd}(k,0;\gamma,\delta) &= \frac{4}{3}k(1-k^2)(2-3k^2)\sqrt{2(1-3k^2)}\,\gamma\\ &\quad + \frac{4}{15}k(1-k^2)(12-2k^2-27k^4)\sqrt{2(1-3k^2)}\,\delta + {\rm O}(2)\\ &= \frac{8\sqrt{2}}{3}k\left[(1+{\rm O}(k))\gamma + \left(\frac{6}{5} + {\rm O}(k)\right)\delta + {\rm O}(2)\right],\end{aligned}$$

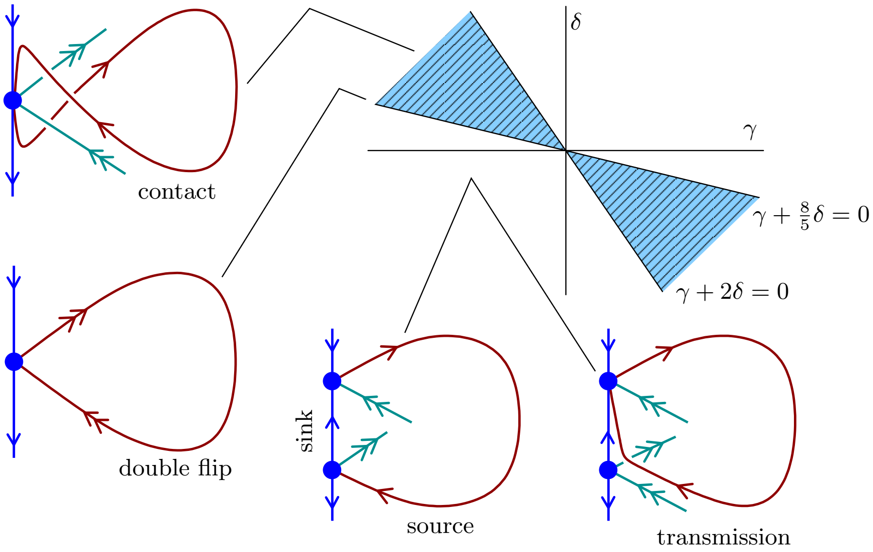


Figure 7.3. *The plot in the upper right corner shows the parameter space* (γ, δ). *Contact defects exist in the shaded region, while sources and transmission defects exist outside. Along the line* $\gamma + \frac{8}{5}\delta = 0$, *contact defects and sources collide in a double-flip configuration as outlined in section 6.4.*

where we exploited the fact that the splitting distance vanishes at $k = 0$ due to reversibility of (7.12). We infer from the above expression that the center-unstable and center-stable manifolds intersect transversely at $k = 0$ provided $\gamma + \frac{6}{5}\delta + \mathrm{O}(2) \neq 0$. This curve, along which transversality fails, lies outside the existence cone (7.31) for contact defects, which proves that the reversible contact defects are robust and persist under variations of k_*.

Theorem 7.1. *Equation (7.31) defines a nonempty open cone in* (γ, δ)*-space with the property that the complex CQGL (7.4) has, for each fixed* (γ, δ) *in that cone, a one-parameter family of elementary transverse contact defects parametrized by the wavenumber* k *with* k *close to zero.*

We remark that, at the boundary $\gamma + \frac{8}{5}\delta = 0$ of (7.31), a double-flip transition between sources and contact defects occurs, while at the other boundary $\gamma + 2\delta = 0$, the nonlinear dispersion relation of the asymptotic wave trains of the contact defects changes from convex to concave.

Transmission defects can be found as bound states of the Nozaki–Bekki holes and the small sinks that arise in the unfolding of the saddle node on the center manifold. We refer to Figure 7.3 for an illustration.

8. Conclusion.

We have presented a framework that allows us to systematically study interfaces between nonlinear wave trains with possibly different wavenumber. We used this approach to analyze sinks, contact defects, transmission defects, and sources which are specific defects with distinguished characteristics that occur robustly in reaction-diffusion systems. In addition, we discussed the stability and interaction properties of these defects and investigated some of their bifurcations.

The major open problem is the completeness of the above classification. Other open issues are the nonlinear stability of contact defects, transmission defects, and sources. We also emphasize that the list of bifurcations that we discussed is not exhaustive. Last, it would be interesting to see how individual defects can be accounted for in a macroscopic description of oscillatory media by coupling mean-field equations of Burgers type to ordinary differential equations for the position and phases of defects.

Appendix A. Invariant manifolds.

We prove the invariant-manifold Theorem 5.1 and establish the existence of exponential dichotomies for the linearization of the spatial-dynamical system about the defect.

A.1. Existence of center-stable manifolds.

Our first goal is to prove the existence of a center-stable manifold for the modulated-wave equation

$$\begin{aligned} u_\xi &= v, \\ v_\xi &= D^{-1}[\omega_\mathrm{d}\partial_\tau u - c_\mathrm{d} v - f(u)] \end{aligned} \tag{A.1}$$

near a given defect. We use the notation $\mathbf{u} = (u, v)$ and recall the spaces

$$Y = H^{1/2}(S^1) \times L^2(S^1), \qquad Y^1 = H^1(S^1) \times H^{1/2}(S^1).$$

Throughout the appendix, we assume that $\mathbf{u}_\mathrm{d}(\xi, \tau)$ is a defect solution of (A.1) so that $\mathbf{u}_\mathrm{d}(\xi, \cdot) - \mathbf{u}_\mathrm{wt}(k_\mathrm{d}\xi + \theta_\mathrm{d}(\xi) - \cdot)$ converges to zero as $\xi \to \infty$ for an appropriate differentiable phase-correction function $\theta_\mathrm{d}(\xi)$ with $\theta_\mathrm{d}'(\xi) \to 0$ as $\xi \to \infty$. We will show later how changes of (ω, c) can be accounted for.

We write (A.1) as the abstract differential equation

$$\mathbf{u}_\xi = \mathcal{A}_0\mathbf{u} + \mathcal{F}(\mathbf{u}), \tag{A.2}$$

where

$$\mathcal{A}_0 = \begin{pmatrix} 0 & 1 \\ \omega_\mathrm{d} D^{-1}\partial_\tau & -c_\mathrm{d} D^{-1} \end{pmatrix}, \qquad \mathcal{F}(\mathbf{u}) = \begin{pmatrix} 0 \\ -D^{-1} f(u) \end{pmatrix}. \tag{A.3}$$

We use the ansatz

$$\mathbf{u}(\xi, \tau) = \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \tau) + \mathbf{w}(\xi, \tau) \tag{A.4}$$

together with the pointwise normalization

$$\langle \mathbf{w}(\xi, \cdot), \partial_\tau \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot)\rangle_Y = 0. \tag{A.5}$$

We emphasize that we do not shift $\mathbf{w}$ relative to the group since this would change the type of the equation we are considering. Equation (A.2) becomes

$$\theta'\partial_\tau \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot) + \mathbf{w}_\xi = \mathcal{A}_0\mathbf{w} + \mathcal{F}(\mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot) + \mathbf{w}) - \mathcal{F}(\mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot)). \tag{A.6}$$

For later use, we differentiate (A.5) and obtain the relation

$$\langle \partial_\xi \mathbf{w}(\xi, \cdot), \partial_\tau \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot)\rangle = -\langle \mathbf{w}(\xi, \cdot), [\partial_\xi + \theta'(\xi)\partial_\tau]\partial_\tau \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot)\rangle. \tag{A.7}$$

From now on, we shall use the notation $\mathbf{u}^\theta(\xi) := \mathbf{u}(\xi, \theta(\xi) + \cdot)$. To derive differential equations for θ and $\mathbf{w}$, we take the scalar product of (A.6) with $\partial_\tau \mathbf{u}_\mathrm{d}(\xi, \theta(\xi) + \cdot)$ and substitute (A.7), which gives the equation

$$\begin{aligned}(\text{A.8})\quad \theta_\xi &= \frac{1}{|\partial_\tau \mathbf{u}_\mathrm{d}|^2 - \langle \mathbf{w}, \partial_{\tau\tau} \mathbf{u}_\mathrm{d}^\theta \rangle} \left(\left\langle \partial_{\tau\xi} \mathbf{u}_\mathrm{d}^\theta, \mathbf{w} \right\rangle + \left\langle \partial_\tau \mathbf{u}_\mathrm{d}^\theta, \mathcal{A}_0 \mathbf{w} + \mathcal{F}(\mathbf{u}_\mathrm{d}^\theta + \mathbf{w}) - \mathcal{F}(\mathbf{u}_\mathrm{d}^\theta) \right\rangle \right) \\ &= \mathcal{B}_0(\theta)\mathbf{w} + \mathcal{G}_0(\theta, \mathbf{w})\end{aligned}$$

for θ. The scalar products and norms that appear in (A.8) are in Y, and we introduced the linear operator

$$(\text{A.9})\qquad \mathcal{B}_0(\theta) := \frac{\langle \partial_{\tau\xi} \mathbf{u}_\mathrm{d}^\theta + [\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}^\theta)]^* \partial_\tau \mathbf{u}_\mathrm{d}^\theta, \cdot \rangle}{|\partial_\tau \mathbf{u}_\mathrm{d}|^2},$$

where * denotes the adjoint, with the nonlinear term $\mathcal{G}_0(\theta, \mathbf{w})$ making up the difference. For $\mathbf{w}$, we obtain

$$\begin{aligned}(\text{A.10})\qquad \mathbf{w}_\xi = \mathcal{A}_0 \mathbf{w} + \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}^\theta)\mathbf{w} &+ \left[\mathcal{F}(\mathbf{u}_\mathrm{d}^\theta + \mathbf{w}) - \mathcal{F}(\mathbf{u}_\mathrm{d}^\theta) - \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}^\theta)\mathbf{w} \right] \\ &- \left[\mathcal{B}_0(\theta)\mathbf{w} + \mathcal{G}_0(\theta, \mathbf{w}) \right] \partial_\tau \mathbf{u}_\mathrm{d}^\theta.\end{aligned}$$

To prepare for the later use of the contraction-mapping principle, we modify the nonlinear terms. We choose two cut-off functions $\chi_0(r)$ and $\chi_1(r)$ such that $\chi_0'(r) \geq 0$, $\chi_1'(r) \leq 0$, and

$$\chi_0(r) = \begin{cases} r & \text{for} \quad 0 \leq r \leq 1, \\ 2 & \text{for} \quad r \geq 2, \end{cases} \qquad \chi_1(r) = \begin{cases} 1 & \text{for} \quad 0 \leq r \leq 1, \\ 0 & \text{for} \quad r \geq 2. \end{cases}$$

Instead of (A.8) and (A.10), we then consider the equations

$$(\text{A.11})\qquad \theta_\xi = \delta \chi_0 \left(\frac{\mathcal{B}_0(\theta)\mathbf{w} + \mathcal{G}_0(\theta, \mathbf{w})}{\delta} \right),$$

$$(\text{A.12})\qquad \mathbf{w}_\xi = \mathcal{A}(\theta)\mathbf{w} + \mathcal{G}_1(\theta, \mathbf{w}),$$

where

$$\begin{aligned}(\text{A.13})\qquad \mathcal{A}(\theta) &= \mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}^\theta) - \left[\partial_\tau \mathbf{u}_\mathrm{d}^\theta \right] \mathcal{B}_0(\theta), \\ \mathcal{G}_1(\theta, \mathbf{w}) &= \chi_1 \left(\frac{|\mathbf{w}|_Y}{\delta} \right) \left[\mathcal{F}(\mathbf{u}_\mathrm{d}^\theta + \mathbf{w}) - \mathcal{F}(\mathbf{u}_\mathrm{d}^\theta) - \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}^\theta)\mathbf{w} - \mathcal{G}_0(\theta, \mathbf{w}) \partial_\tau \mathbf{u}_\mathrm{d}^\theta \right].\end{aligned}$$

We emphasize that our cut-off preserves the S^1-equivariance with respect to the symmetry $\theta \mapsto \theta + \check{\theta}$ for $\check{\theta} \in \mathbb{R}$. In addition, the cut-off is chosen so that solutions to the modified equations (A.11)–(A.12) give solutions to the original equations (A.8)–(A.10) provided $|\mathbf{w}|_Y$ is less than δ for all ξ. In other words, there is no restriction on the norm of θ.

We remark that

$$\begin{aligned} |\mathcal{G}_0(\theta, \mathbf{w})| &= \mathrm{O}(|\mathbf{w}|_Y^2), \qquad & |\partial_{(\theta,\mathbf{w})} \mathcal{G}_0(\theta, \mathbf{w})|_{L(\mathbb{R} \times Y^1, \mathbb{R})} &= \mathrm{O}(|\mathbf{w}|_Y), \\ |\mathcal{G}_1(\theta, \mathbf{w})|_Y &= \mathrm{O}(|\mathbf{w}|_{Y^1}^2), \qquad & |\partial_{(\theta,\mathbf{w})} \mathcal{G}_1(\theta, \mathbf{w})|_{L(\mathbb{R} \times Y^1, Y)} &= \mathrm{O}(|\mathbf{w}|_{Y^1}) \end{aligned}$$

uniformly in θ, which allows us to replace $\mathbf{w} \to \delta\mathbf{w}$ for any small $\delta > 0$. We obtain the new equation

$$\theta_\xi = \delta\chi_0(|\mathbf{w}|_Y)\left[\mathcal{B}_0(\theta)\mathbf{w} + \frac{1}{\delta}\mathcal{G}_0(\theta,\delta\mathbf{w})\right] \;=\; \delta\check{\mathcal{G}}_0(\theta,\mathbf{w}), \tag{A.14}$$

$$\mathbf{w}_\xi = \mathcal{A}(\theta)\mathbf{w} + \frac{1}{\delta}\mathcal{G}_1(\theta,\delta\mathbf{w}) \;=\; \mathcal{A}(\theta)\mathbf{w} + \check{\mathcal{G}}_1(\theta,\mathbf{w}). \tag{A.15}$$

There is a constant $C_0 > 0$ so that $\check{\mathcal{G}}_0$ and $\check{\mathcal{G}}_1$ are bounded by C_0 with Lipschitz constants less than $C_0\delta$ uniformly in $(\theta,\mathbf{w})$.

Lemma A.1. *There are positive numbers κ and C_1 such that the equation*

$$\mathbf{w}_\xi = \mathcal{A}(\theta_0)\mathbf{w} \tag{A.16}$$

has exponential dichotomies $\Phi^{\mathrm{cs}}_{\theta_0}(\xi,\zeta)$ and $\Phi^{\mathrm{uu}}_{\theta_0}(\xi,\zeta)$ with rate κ and constant C_1 for each $\theta_0 \in \mathbb{R}$.

Proof. We will prove the lemma by beginning with an equation for which we know that exponential dichotomies exist and then successively changing the equation until we arrive at (A.16). We will make sure that the equations will have exponential dichotomies at each stage by appealing to the roughness theorem for dichotomies [39] and to Theorem A.4 in section A.2. To save notation, we consider the linear equation

$$\mathbf{w}_\xi = \mathcal{A}\mathbf{w}$$

and simply give the different operators $\mathcal{A}$ in our chain of equations.

We begin by recalling that the linearization $\mathcal{A} = \mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_{\mathrm{wt}}(k\xi+\theta_0))$ about the asymptotic wave train has center-stable and strong-unstable exponential dichotomies uniformly in θ_0 by [34, Chapter 2]. Since $\theta_{\mathrm{d}}'(\xi) \to 0$ as $\xi \to \infty$, Theorem A.4 in section A.2 implies that there is a $\xi_0 > 0$ so that $\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_{\mathrm{wt}}(k\xi + \theta_{\mathrm{d}}(\xi)))$ has center-stable and strong-unstable exponential dichotomies for $\xi \geq \xi_0$. Invoking the roughness theorem [39, Theorem 1] finally shows that $\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_{\mathrm{d}}(\xi))$ has center-stable and strong-unstable dichotomies for $\xi \geq 0$ with rate κ and constant C_1.

In summary, we showed that

$$\mathbf{w}_\xi = [\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_{\mathrm{d}}(\xi))]\,\mathbf{w} \tag{A.17}$$

has center-stable and strong-unstable dichotomies for $\xi \geq 0$. It therefore remains to establish the same result for

$$\mathbf{w}_\xi = [\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_{\mathrm{d}}(\xi)) - [\partial_\tau \mathbf{u}_{\mathrm{d}}]\,\mathcal{B}_0(0)]\,\mathbf{w}, \tag{A.18}$$

where $\theta \equiv 0$. Note that, by construction, any solution $\mathbf{w}(\xi)$ of (A.18) for which $\langle \mathbf{w}(0), \partial_\tau \mathbf{u}_{\mathrm{d}}(0)\rangle = 0$ satisfies

$$\langle \mathbf{w}(\xi), \partial_\tau \mathbf{u}_{\mathrm{d}}(\xi)\rangle = 0 \tag{A.19}$$

for all ξ. It is then not difficult to see that (A.18) restricted to solutions that satisfy (A.19) has center-stable and strong-unstable dichotomies since (A.18) is merely part of the decomposition of solutions to (A.17) into $\mathbf{w}(\xi)$ and $\alpha(\xi)\partial_\tau \mathbf{u}_\mathrm{d}(\xi)$ for real-valued functions α. Since we wish to consider (A.18) for all $\mathbf{w}$, we have to calculate the solution to (A.18) with initial data $\mathbf{w}(0) = \partial_\tau \mathbf{u}_\mathrm{d}(0)$. Fortunately, it is easy to see that $\alpha(\xi)\partial_\tau \mathbf{u}_\mathrm{d}(\xi)$ is the desired solution to (A.18), where $\alpha(\xi)$ satisfies

$$\alpha_\xi = -\alpha \frac{\mathrm{d}}{\mathrm{d}\xi}\left[\log |\partial_\tau \mathbf{u}_\mathrm{d}(\xi)|_Y^2\right], \qquad \alpha(0) = 1,$$

so that $\alpha(\xi) = |\partial_\tau \mathbf{u}_\mathrm{d}(0)|_Y^2 / |\partial_\tau \mathbf{u}_\mathrm{d}(\xi)|_Y^2$ for all ξ. This shows that (A.18) with $\mathbf{w} \in Y$ has center-stable and strong-unstable dichotomies for $\xi \geq 0$. The remaining statements are once more a consequence of Theorem A.4. ■

The following corollary summarizes the first part of the proof of the preceding lemma.

Corollary A.2. *The linearization*

$$\mathbf{w}_\xi = [\mathcal{A}_0 + \mathcal{F}_\mathbf{u}(\mathbf{u}_\mathrm{d}(\xi))]\,\mathbf{w}$$

about the defect has exponential dichotomies in appropriate weighted spaces if and only if the linearization about the asymptotic wave trains has dichotomies in these spaces.

For each $\eta \in \mathbb{R}$, we define the spaces

$$\mathcal{Y}_\eta = L^2_\eta(\mathbb{R}^+, Y), \qquad \mathcal{Y}^1_\eta = L^2_\eta(\mathbb{R}^+, Y^1) \cap H^1_\eta(\mathbb{R}^+, Y), \tag{A.20}$$

where L^2_η refers to the exponentially weighted L^2-space with norm

$$|\mathbf{u}|_{L^2_\eta(\mathbb{R}^+)} = |\mathrm{e}^{-\eta\xi}\mathbf{u}(\xi)|_{L^2(\mathbb{R}^+)}.$$

We are interested in proving the existence of a center-stable manifold of class $\mathcal{C}^\ell$, say, for some integer $\ell > 0$ that we fix from now on. We will then choose η_0 and δ with $0 < \delta < \eta_0 < \kappa/\ell$, where κ appeared in Lemma A.1.

For every $\eta > \delta$ and each $\mathbf{w} \in \mathcal{Y}^1_\eta$, (A.14)

$$\theta_\xi = \delta \check{\mathcal{G}}_0(\theta, \mathbf{w}(\xi)), \qquad \theta(0) = 0,$$

can be solved uniquely for $\theta(\xi)$, and we have $\sup_{\xi \geq 0} |\theta'(\xi)| \leq C_0\delta$. Furthermore, the function $\mathbf{w} \mapsto \theta$ is Lipschitz as a map from $\mathcal{Y}^1_\eta$ into itself with Lipschitz constant bounded by $C_0\delta/(\eta - \delta)$ (see [42, Lemma 2.4(ii)]). We denote this map by $\theta = \theta[\mathbf{w}]$.

Thus, it remains to solve (A.15) for $\mathbf{w}$, which reads

$$\mathbf{w}_\xi = \mathcal{A}(\theta[\mathbf{w}])\mathbf{w} + \check{\mathcal{G}}_1(\theta[\mathbf{w}], \mathbf{w}) \tag{A.21}$$

after substituting $\theta = \theta[\mathbf{w}]$. After decreasing $\delta > 0$ if necessary, Lemma A.1 together with the properties of the map $\mathbf{w} \mapsto \theta[\mathbf{w}]$ imply that the linear part of (A.21) has center-stable and strong-unstable dichotomies $\Phi^\mathrm{cs}_{\theta[\mathbf{w}]}(\xi, \zeta)$ and $\Phi^\mathrm{uu}_{\theta[\mathbf{w}]}(\xi, \zeta)$, respectively, which allows us to write (A.21) as the integral equation

$$\begin{aligned}\mathbf{w}(\xi) = \Phi^\mathrm{cs}_{\theta[\mathbf{w}]}(\xi, 0)\mathbf{w}^\mathrm{cs}_0 &+ \int_0^\xi \Phi^\mathrm{cs}_{\theta[\mathbf{w}]}(\xi, \zeta)\check{\mathcal{G}}_1(\theta[\mathbf{w}](\zeta), \mathbf{w}(\zeta))\,\mathrm{d}\zeta \\ &+ \int_\infty^\xi \Phi^\mathrm{uu}_{\theta[\mathbf{w}]}(\xi, \zeta)\check{\mathcal{G}}_1(\theta[\mathbf{w}](\zeta), \mathbf{w}(\zeta))\,\mathrm{d}\zeta.\end{aligned} \tag{A.22}$$

From this point on, we can proceed, with very minor modifications, as in the proof of Fenichel's theorem given in [42, section 2 and Appendix A] by setting up a contraction-mapping argument on the space $\mathcal{Y}^1_{\eta_0}$. Since the proof in [42] is very detailed, we decided not to repeat it here. We should, however, comment on an additional result that we need to make the proof in [42] work. In order to apply the arguments on [42, page 73], we need the following optimal-regularity result.

Lemma A.3. *If we fix any function $\theta(\xi)$ with $\sup_{\xi\geq 0}|\theta'(\xi)|\leq C_0\delta$ and denote by $\Phi^{\mathrm{cs}}_\theta(\xi,\zeta)$ and $\Phi^{\mathrm{uu}}_\theta(\xi,\zeta)$ the dichotomies of*

$$\mathbf{w}_\xi = \mathcal{A}(\theta(\xi))\mathbf{w},$$

then the operator $\mathcal{S}:\mathcal{Y}_\eta\to\mathcal{Y}^1_\eta$ defined by

$$[\mathcal{S}\mathbf{g}](\xi)=\int_0^\xi \Phi^{\mathrm{cs}}_\theta(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta+\int_\infty^\xi \Phi^{\mathrm{uu}}_\theta(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta$$

is well defined and bounded as long as $0<\eta<\kappa$.

Proof. Define $E^{\mathrm{u}}_0=\mathrm{Rg}(\Phi^{\mathrm{uu}}_\theta(0,0))$, and consider the operator

$$\mathcal{T}:\quad \mathrm{D}(\mathcal{T})\longrightarrow\mathcal{Y}_\eta,\quad \mathbf{u}\longmapsto\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}\xi}-A(\theta(\cdot))\mathbf{w} \tag{A.23}$$

with domain

$$\mathrm{D}(\mathcal{T})=\mathcal{Y}^1_\eta\cap\{\mathbf{w}\in\mathcal{Y}^1_\eta;\ \mathbf{w}(0)\in E^{\mathrm{u}}_0\}. \tag{A.24}$$

It follows as in [49, section 5.2] that $\mathcal{T}$ is a closed unbounded operator on $\mathcal{Y}_\eta$ with inverse given by $\mathcal{S}$, which proves the lemma. ■

The optimal-regularity result allows us also to prove smooth dependence of the fixed point on the parameters (ω,c). Indeed, if we define

$$\check{\mathcal{F}}(\check{\mathbf{u}})=\begin{pmatrix} v\\ -\check{\omega}D^{-1}\partial_\tau u-\check{c}D^{-1}v-D^{-1}f(u)\end{pmatrix},\qquad \check{\mathbf{u}}=(u,v,\check{\omega},\check{c}),$$

we can replace $\mathbf{u}$ and $\mathcal{F}$ by $\check{\mathbf{u}}$ and $\check{\mathcal{F}}$ and proceed exactly as above.

Finally, we mention that the resulting center-stable manifold is constructed in the artificially augmented phase space $\mathbf{w}\in Y^1$. The intersection of this manifold with the smooth bundle defined by (A.5) is transverse, since the normal direction to the bundle is contained in the center subspace. Therefore, the "true" center-stable manifold, defined as the intersection of the center-stable manifold that we constructed above and the bundle given by (A.5), is the desired smooth and flow-invariant center-stable manifold. The statements about fibers can be proved as in [42].

A.2. Slowly varying coefficients. We consider the equation

$$\mathbf{w}_\xi=\mathcal{A}(\theta)\mathbf{w} \tag{A.25}$$

for a given function $\theta = \theta(\xi)$, where $A(\theta)$ has been defined in (A.13). We assume that (A.25) has an exponential dichotomy on $\mathbb{R}^+$ for each constant function $\theta(\xi) \equiv \theta_0 \in \mathbb{R}$ and that the rate κ and the constant C of the dichotomies can be chosen independently of $\theta_0 \in \mathbb{R}$.

The following theorem, proved in [7, 24] for ODEs and PDEs, respectively, shows that (A.25) then has dichotomies for any slowly varying function $\theta(\xi)$.

Theorem A.4. *There are positive constants ε_0, $\check{\kappa}$, and $\check{C}$ such that* (A.25) *has an exponential dichotomy on $\mathbb{R}^+$ with rate $\check{\kappa}$ and constant $\check{C}$ for each differentiable function $\theta(\xi)$ with $\theta(0) = 0$ and $\sup_{\xi \geq 0} |\theta'(\xi)| \leq \varepsilon_0$.*

Proof. We denote by $\Phi^{\mathrm{s}}_{\theta_0}(\xi,\zeta)$ and $\Phi^{\mathrm{u}}_{\theta_0}(\xi,\zeta)$ the exponential dichotomies of the equation

$$\mathbf{w}_\xi = \mathcal{A}(\theta_0)\mathbf{w}$$

for constant functions $\theta(\xi) \equiv \theta_0 \in \mathbb{R}$. Since the θ_0-dependent terms of the operator $\mathcal{A}(\theta_0)$ are bounded and depend smoothly on θ_0 (see (A.13) and (A.9)), the robustness theorem for exponential dichotomies [7, 24, 39] implies that these dichotomies can be chosen to depend smoothly on $\theta_0 \in \mathbb{R}$. In fact, there are positive constants C and κ such that

$$\left\| \frac{\mathrm{d}\Phi^{\mathrm{s}}_{\theta_0}}{\mathrm{d}\theta}(\xi,\zeta) \right\| + \left\| \frac{\mathrm{d}\Phi^{\mathrm{u}}_{\theta_0}}{\mathrm{d}\theta}(\zeta,\xi) \right\| \leq C\mathrm{e}^{-\kappa|\xi-\zeta|} \tag{A.26}$$

uniformly in $\xi \geq \zeta \geq 0$. We denote by E^{u}_0 the range of the unstable projection $\Phi^{\mathrm{u}}_0(0,0)$ for $\theta_0 = 0$.

Next, pick a function $\theta(\xi)$ so that $\theta(0) = 0$ and $\sup_{\xi \geq 0} |\theta'(\xi)| \leq \varepsilon_0$. Consider the operator

$$\mathcal{T}(\theta): \quad \mathrm{D}(\mathcal{T}(\theta)) \longrightarrow L^2(\mathbb{R}^+, Y), \quad \mathbf{u} \longmapsto \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\xi} - A(\theta(\cdot))\mathbf{u} \tag{A.27}$$

with domain

$$\mathrm{D}(\mathcal{T}(\theta)) = L^2(\mathbb{R}^+, Y^1) \cap \{\mathbf{u} \in H^1(\mathbb{R}^+, Y);\ \mathbf{u}(0) \in E^{\mathrm{u}}_0\}, \tag{A.28}$$

which is independent of θ. We may consider $\mathcal{T}$ as a closed unbounded operator on $L^2(\mathbb{R}^+, Y)$. If we can prove that $\mathcal{T}(\theta)$ has a bounded inverse with bounds that are independent of θ, then the theorem is proved as we may then proceed as in [49, section 5.3.1] to construct exponential dichotomies of (A.25) with constants and rates that do not depend on θ.

Thus, we have to solve the equation

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\xi} = A(\theta(\xi))\mathbf{u} + \mathbf{g}(\xi) \tag{A.29}$$

for a given $\mathbf{g} \in L^2(\mathbb{R}^+, Y)$. We define

$$\mathbf{u}(\xi) = \left[\check{\mathcal{L}}(\theta)\mathbf{g}\right](\xi) = \int_0^\xi \Phi^{\mathrm{s}}_{\theta(\xi)}(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta + \int_\infty^\xi \Phi^{\mathrm{u}}_{\theta(\xi)}(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta \tag{A.30}$$

and set $\check{\mathcal{L}}(\theta)\mathbf{g} := \mathbf{u}$. It follows from [49] that $\mathbf{u}(\xi)$ lies in $\mathrm{D}(\mathcal{T}(\theta))$ and that

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\xi} = A(\theta(\xi))\mathbf{u} + \mathbf{g}(\xi) + [\mathcal{S}(\theta)\mathbf{g}](\xi), \tag{A.31}$$

where

$$[\mathcal{S}(\theta)\mathbf{g}](\xi) = \left[\int_0^\xi \frac{\mathrm{d}\Phi^{\mathrm{s}}_{\theta(\xi)}}{\mathrm{d}\theta}(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta + \int_\infty^\xi \frac{\mathrm{d}\Phi^{\mathrm{u}}_{\theta(\xi)}}{\mathrm{d}\theta}(\xi,\zeta)\mathbf{g}(\zeta)\,\mathrm{d}\zeta\right]\theta'(\xi). \tag{A.32}$$

Using (A.26), we see that

$$\mathcal{S}(\theta): \quad L^2(\mathbb{R}^+,Y) \longrightarrow L^2(\mathbb{R}^+,Y)$$

with $\|\mathcal{S}(\theta)\| \le \varepsilon_0 C_1$ for some $C_1 > 0$ that does not depend on θ. Therefore, $1+\mathcal{S}(\theta)$ is invertible on $L^2(\mathbb{R}^+,Y)$ uniformly in θ provided $\varepsilon_0 > 0$ is smaller than $2/C_1$. The desired solution to (A.29) is then given by

$$\mathbf{u} := \mathcal{L}(\theta)\mathbf{g} = \check{\mathcal{L}}(\theta)[1+\mathcal{S}(\theta)]^{-1}\mathbf{g},$$

where $\check{\mathcal{L}}(\theta)$ has been defined in (A.30). This proves that $\mathcal{T}(\theta)$ has a bounded inverse on $L^2(\mathbb{R}^+,Y)$ and, together with the fact that $\mathcal{T}(\theta)$ is closed, shows that the inverse is bounded as an operator into the domain of $\mathcal{T}(\theta)$. Since the bounds are clearly independent of θ, we have proved our claim and therefore the theorem. ■

Appendix B. Parameter values for the numerical simulations. The numerical simulations in Figure 1.2(i) and (ii) show the v-component of solutions to the Brusselator

$$\begin{aligned} u_t &= d_1 u_{xx} + a + (b+1)u + u^2 v, \\ v_t &= d_2 v_{xx} + bu - u^2 v \end{aligned} \tag{B.1}$$

on the interval $[0,L]$ with Neumann boundary conditions. The parameters are as in [38] so that

$$d_1 = 4.11, \qquad d_2 = 9.73, \qquad a = 2.5, \qquad b = 10.0.$$

The length L of the spatial interval is chosen between 250 and 650. Sources and sinks are created from the initial condition

$$(u_0,v_0)(x) = \left(a + \frac{1}{2}\tanh\left(x - \frac{L}{2}\right), \frac{b}{a} + \frac{1}{10}\cos x\right),$$

which excites both Turing and Hopf modes. We solved the system (B.1) using a forward Euler scheme in time and centered finite differences in space.

REFERENCES

[1] R. Alvarez, M. van Hecke, and W. van Saarloos, *Sources and sinks separating domains of left- and right-travelling waves: Experiment versus amplitude equations*, Phys. Rev. E (3), 56 (1997), pp. R1306–1309.

[2] I. S. Aranson and L. Kramer, *The world of the complex Ginzburg–Landau equation*, Rev. Modern Phys., 74 (2002), pp. 99–143.

[3] D. Barkley, *A model for fast computer-simulation of waves in excitable media*, Phys. D, 49 (1991), pp. 61–70.

[4] J. Burguete, H. Chaté, F. Daviaud, and N. Mukolobwiez, *Bekki–Nozaki amplitude holes in hydrothermal nonlinear waves*, Phys. Rev. Lett., 82 (1999), pp. 3252–3255.

[5] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, World Scientific, Singapore, 2000.

[6] S.-N. Chow and X.-B. Lin, *Bifurcation of a homoclinic orbit with a saddle-node equilibrium*, Differential Integral Equations, 3 (1990), pp. 435–466.

[7] W. A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, 1978.

[8] B. Dionne, M. Golubitsky, M. Silber, and I. Stewart, *Time-periodic spatially periodic planforms in Euclidean equivariant partial differential equations*, Philos. Trans. Roy. Soc. London Ser. A, 352 (1995), pp. 125–168.

[9] E. Doedel, A. R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. Sandstede, and X. Wang, AUTO97: *Continuation and Bifurcation Software for Ordinary Differential Equations (with* HOMCONT*)*, Technical report, Concordia University, Montreal, Quebec, Canada, 1997.

[10] A. Doelman, *Breaking the hidden symmetry in the Ginzburg–Landau equation*, Phys. D, 97 (1996), pp. 398–428.

[11] A. Doelman, B. Sandstede, A. Scheel, and G. Schneider, *The Dynamics of Modulated Wave Trains*, in preparation.

[12] W. Eckhaus and G. Iooss, *Strong selection or rejection of spatially periodic patterns in degenerate bifurcations*, Phys. D, 39 (1989), pp. 124–146.

[13] S.-I. Ei, *The motion of weakly interacting pulses in reaction-diffusion systems*, J. Dynam. Differential Equations, 14 (2002), pp. 85–137.

[14] C. Elphick, E. Meron, and E. A. Spiegel, *Patterns of propagating pulses*, SIAM J. Appl. Math., 50 (1990), pp. 490–503.

[15] H. Engler, *Asymptotic stability of travelling wave solutions for perturbations with algebraic decay*, J. Differential Equations, 185 (2002), pp. 348–369.

[16] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.

[17] T. Gallay, G. Schneider, and H. Uecker, *Stable transport of information near essentially unstable localized structures*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 349–390.

[18] A. Goryachev, H. Chaté, and R. Kapral, *Synchronization defects and broken symmetry in spiral waves*, Phys. Rev. Lett., 80 (1998), pp. 873–876.

[19] P. Habdas, M. J. Case, and J. R. de Bruyn, *Behavior of sinks and source defects in a one-dimensional traveling finger pattern*, Phys. Rev. E (3), 63 (2001), 066305.

[20] C. T. Hamik and O. Steinbock, *Shock structures and bunching fronts in excitable reaction-diffusion systems*, Phys. Rev. E (3), 65 (2002), 046224.

[21] M. van Hecke, *Building blocks of spatiotemporal intermittency*, Phys. Rev. Lett., 80 (1998), pp. 1896–1899.

[22] M. van Hecke, *Coherent and incoherent structures in systems described by the* 1*D CGLE: Experiments and identification*, Phys. D, 174 (2003), pp. 134–151.

[23] M. van Hecke, C. Storm, and W. van Saarloos, *Sources, sinks and wavenumber selection in coupled CGL equations and experimental implications for counter-propagating wave systems*, Phys. D, 134 (1999), pp. 1–47.

[24] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 804, Springer-Verlag, New York, 1981.

[25] P. Howard, *Pointwise estimates and stability for degenerate viscous shock waves*, J. Reine Angew. Math., 545 (2002), pp. 19–65.

[26] L. N. Howard and N. Kopell, *Slowly varying waves and shock structures in reaction-diffusion equations*, Stud. Appl. Math., 56 (1976/77), pp. 95–145.

[27] G. Iooss, A. Mielke, and Y. Demay, *Theory of steady Ginzburg–Landau equation, in hydrodynamic stability problems*, Eur. J. Mech. B Fluids, 8 (1989), pp. 229–268.

[28] T. Kapitula, *Stability of weak shocks in λ-ω systems*, Indiana Univ. Math. J., 40 (1991), pp. 1193–1219.

[29] T. Kapitula, *Existence and stability of singular heteroclinic orbits for the Ginzburg–Landau equation*, Nonlinearity, 9 (1996), pp. 669–685.

[30] K. Kirchgässner, *Wave-solutions of reversible systems and applications*, J. Differential Equations, 45 (1982), pp. 113–127.

[31] P. Kolodner, *Extended states of nonlinear traveling-wave convection* II: *Fronts and spatiotemporal defects*, Phys. Rev. A (3), 45 (1992), pp. 6452–6468.

[32] N. Kopell and L. N. Howard, *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52 (1973), pp. 291–328.

[33] G. J. Lord, D. Peterhof, B. Sandstede, and A. Scheel, *Numerical computation of solitary waves in infinite cylindrical domains*, SIAM J. Numer. Anal., 37 (2000), pp. 1420–1454.

[34] A. Mielke, *A spatial center manifold approach to steady bifurcations from spatially periodic patterns*, in Dynamics of Nonlinear Waves in Dissipative Systems: Reductions, Bifurcations and Stability, G. Dangelmayr, B. Fiedler, K. Kirchgässner, and A. Mielke, eds., Pitman Res. Notes Math. Ser. 352, Longman, Harlow, UK, 1996.

[35] M. Murata, *Large time asymptotics for fundamental solutions of diffusion equations*, Tohoku Math. J. (2), 37 (1985), pp. 151–195.

[36] L. Pastur, M.-T. Westra, and W. van de Water, *Sources and sinks in* 1*D traveling waves*, Phys. D, 174 (2003), pp. 71–83.

[37] L. Pastur, M.-T. Westra, D. Snouck, W. van de Water, M. van Hecke, C. Storm, and W. van Saarloos, *Sources and holes in a one-dimensional traveling-wave convection experiment*, Phys. Rev. E (3), 67 (2003), 036305.

[38] J.-J. Perraud, A. De Wit, E. Dulos, P. De Kepper, G. Dewel, and P. Borckmans, *One-dimensional "spirals": Novel asynchronous chemical wave sources*, Phys. Rev. Lett., 71 (1993), pp. 1272–1275.

[39] D. Peterhof, B. Sandstede, and A. Scheel, *Exponential dichotomies for solitary-wave solutions of semilinear elliptic equations on infinite cylinders*, J. Differential Equations, 140 (1997), pp. 266–308.

[40] J. Rademacher, *Tracefiring of Pulses*, in preparation.

[41] W. van Saarloos and P. C. Hohenberg, *Fronts, pulses, sources and sinks in generalized complex Ginzburg–Landau equations*, Phys. D, 56 (1992), pp. 303–367.

[42] K. Sakamoto, *Invariant manifolds in singular perturbation problems for ordinary differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 45–78.

[43] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II, B. Fiedler, ed., Elsevier, Amsterdam, 2002, pp. 983–1055.

[44] B. Sandstede and A. Scheel, *Essential instability of pulses, and bifurcations to modulated travelling waves*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 1263–1290.

[45] B. Sandstede and A. Scheel, *Spectral stability of modulated travelling waves bifurcating near essential instabilities*, Proc. Roy. Soc. Edinburgh Sect. A, 130 (2000), pp. 419–448.

[46] B. Sandstede and A. Scheel, *Absolute versus convective instability of spiral waves*, Phys. Rev. E (3), 62 (2000), pp. 7708–7714.

[47] B. Sandstede and A. Scheel, *Absolute and convective instabilities of waves on unbounded and large bounded domains*, Phys. D, 145 (2000), pp. 233–277.

[48] B. Sandstede and A. Scheel, *Essential instabilities of fronts: Bifurcation, and bifurcation failure*, Dyn. Syst., 16 (2001), pp. 1–28.

[49] B. Sandstede and A. Scheel, *On the structure of spectra of modulated travelling waves*, Math. Nachr., 232 (2001), pp. 39–93.

[50] B. Sandstede and A. Scheel, *Evans function and blow-up methods in critical eigenvalue problems*, Discrete Contin. Dyn. Syst. Ser. B, to appear.

[51] B. Sandstede and A. Scheel, *Absolute Instabilities of Pulses*, preprint, The Ohio State University, Columbus, OH, 2003.

[52] B. Sandstede and A. Scheel, *Period-Doubling Bifurcations of Spiral Waves*, in preparation.

[53] B. Sandstede and A. Scheel, *Stability and Instability of Spiral Waves*, in preparation.

[54] D. H. Sattinger, *On the stability of waves of nonlinear parabolic systems*, Adv. Math., 22 (1976), pp. 312–355.

[55] A. Scheel, *Radially symmetric patterns of reaction-diffusion systems*, Mem. Amer. Math. Soc., 165 (2003).

[56] G. Schneider, *Hopf bifurcation in spatially extended reaction-diffusion systems*, J. Nonlinear Sci., 8 (1998), pp. 17–41.

[57] A. Vanderbauwhede and B. Fiedler, *Homoclinic period blow-up in reversible and conservative systems*, Z. Angew. Math. Phys., 43 (1992), pp. 292–318.

[58] M. Yoneyama, A. Fujii, and S. Maeda, *Wavelength-doubled spiral fragments in photosensitive monolayers*, J. Amer. Chem. Soc., 117 (1995), pp. 8188–9191.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 1, pp. 69–83

Hamiltonian Particle-Mesh Method for Two-Layer Shallow-Water Equations Subject to the Rigid-Lid Approximation*

Colin J. Cotter[†], Jason Frank[‡], and Sebastian Reich[†]

Abstract. We develop a particle-mesh method for two-layer shallow-water equations subject to the rigid-lid approximation. The method is based on the recently proposed Hamiltonian particle-mesh (HPM) method and the interpretation of the rigid-lid approximation as a set of holonomic constraints. The suggested spatial discretization leads to a constrained Hamiltonian system of ODEs which is integrated in time using a variant of the symplectic SHAKE/RATTLE algorithm. It is demonstrated that the elimination of external gravity waves by the rigid-lid approximation can be achieved in a computationally stable and efficient way.

Key words. particle methods, rotating shallow-water equations, symplectic integration, rigid-lid approximation

AMS subject classifications. 76M28, 37M15, 65P10, 76U05, 76B60, 70-08, 70H05, 70F20

DOI. 10.1137/030600076

1. Introduction.

Theorists frequently regard the ocean as a two-layer fluid with the interface between layers corresponding to the main thermocline. This idealization is perhaps most appropriate in the northwestern subtropical North Atlantic. Consider, then, a rotating fluid composed of two immiscible layers with different constant densities $\rho_1 < \rho_2$ over a flat bottom topography at $z = 0$. See Figure 1.1 and the excellent exposition [21]. Under the assumption that $\rho_1 \approx \rho_2$, the associated two-layer shallow-water equations are

$$\frac{D}{Dt_i}\mathbf{u}_i + f\mathbf{u}_i^{\perp} = \begin{cases} -g\nabla_{\mathbf{x}}(h_1 + h_2), & i = 1, \\ -g\nabla_{\mathbf{x}}(h_1 + h_2) - g'\nabla_{\mathbf{x}}h_2, & i = 2, \end{cases} \tag{1.1}$$

where $\mathbf{u}_i \equiv (u_i, v_i)^T$ is the horizontal velocity in the ith layer, $f > 0$ is the *Coriolis parameter*, $\mathbf{u}_i^{\perp} \equiv (-v_i, u_i)^T$,

$$\frac{D}{Dt_i} \equiv \frac{\partial}{\partial t} + \mathbf{u}_i \cdot \nabla_{\mathbf{x}}, \qquad \text{and} \qquad g' \equiv \frac{\rho_2 - \rho_1}{\rho_2} g$$

[*] Received by the editors March 3, 2003; accepted for publication (in revised form) by D. Holm November 21, 2003; published electronically March 4, 2004.

http://www.siam.org/journals/siads/3-1/60007.html

[†] Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom (colin.cotter@imperial.ac.uk, s.reich@imperial.ac.uk).

[‡] CWI, P. O. Box 94079, 1090 GB Amsterdam, The Netherlands (jason@cwi.nl).

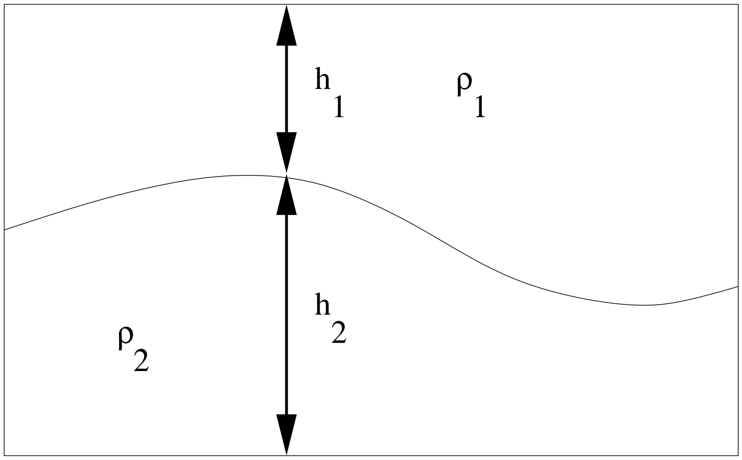


Figure 1.1. *Two-layer shallow-water model with rigid lid.*

is the *reduced gravity.*[1] By assumption, $g' \ll g$. Each layer-depth h_i satisfies the continuity equation

$$\frac{\partial h_i}{\partial t} + \nabla_{\mathbf{x}} \cdot (h_i \mathbf{u}_i) = 0. \tag{1.2}$$

It is also reasonable to assume that the combined flow in both layers is incompressible, which leads to the *rigid-lid* constraint

$$h \equiv h_1 + h_2 = H = \text{const.} \tag{1.3}$$

Equation (1.1) is replaced by

$$\frac{D}{Dt_i}\mathbf{u}_i + f\mathbf{u}_i^{\perp} = \begin{cases} -\nabla_{\mathbf{x}} p, & i = 1, \\ -\nabla_{\mathbf{x}} p - g'\nabla_{\mathbf{x}} h_2, & i = 2, \end{cases}$$

where p is the pressure field enforcing the rigid-lid constraint (1.3) which, after differentiation in time, is equivalent to

$$\nabla_{\mathbf{x}} \cdot (h_1 \mathbf{u}_1) + \nabla_{\mathbf{x}} \cdot (h_2 \mathbf{u}_2) = 0.$$

We also make the simplifying assumption that both layers have a (nondimensionalized) *mean layer-depth* of $H_i = 1$, i.e., $H = H_1 + H_2 = 2$, and replace reduced gravity g' with an appropriate constant c_0.

In a Lagrangian description of the model, we introduce a continuum of fluid particles $\mathbf{X}_i(\mathbf{a}_i, t) \equiv (X_i(\mathbf{a}_i, t), Y_i(\mathbf{a}_i, t))^T$ in each layer $i = 1, 2$, which are *labeled/marked* by their

[1] Equation (1.1) is a slight variation of the formulation given in [21, p. 85]. While (1.1) leads to a Hamiltonian formulation, no obvious Hamiltonian interpretation of (12.3) in [21] could be found. However, both formulations are identical under the rigid-lid approximation.

initial positions $\mathbf{a}_i = \mathbf{X}_i(\mathbf{a}_i, 0)$. Hence the independent variables are time t and labels $\mathbf{a}_i$. The material time derivative D/Dt becomes a partial derivative which, with a slight abuse of notation, we denote by d/dt.

Let $h_i^o(\mathbf{a}_i)$ denote the initial layer-depth at $t = 0$. Then the layer-depth is given at any time t by

$$h_i(\mathbf{x}, t) = \int h_i^o(\mathbf{a})\, \delta(\mathbf{x} - \mathbf{X}_i(\mathbf{a}_i, t))\, d^2\mathbf{a}_i, \qquad i = 1, 2, \tag{1.4}$$

where δ denotes the Dirac delta function. This formula and

$$\frac{d}{dt}\mathbf{X}_i = \mathbf{u}_i$$

replace the continuity equation (1.2) in a Lagrangian description of fluid dynamics. Hence we finally obtain the constrained infinite-dimensional Newtonian equations of motion

$$\begin{aligned}
\frac{d}{dt}\mathbf{u}_i &= -f\mathbf{u}_i^{\perp} - \begin{cases} \nabla_{\mathbf{X}_1} p, & i = 1, \\ \nabla_{\mathbf{X}_2} p + c_0 \nabla_{\mathbf{X}_2} h_2, & i = 2, \end{cases} \\
\frac{d}{dt}\mathbf{X}_i &= \mathbf{u}_i, \\
0 &= h_1(\mathbf{x}, t) + h_2(\mathbf{x}, t) - H.
\end{aligned}$$

In the following section we describe a spatial discretization for this model.

2. The Hamiltonian particle-mesh (HPM) method with rigid-lid constraint.

To simplify the discussion, we assume a double periodic domain $\mathbf{x} \in \mathcal{R} \equiv [-\pi, +\pi)^2$ and introduce a regular grid $\mathbf{x}^{pq}$ on $\mathcal{R}$ with equal grid spacing Δx in the x- and y-direction. Let $\psi^{pq}(\mathbf{x})$ denote the tensor product cubic B-spline centered at $\mathbf{x}^{pq} \equiv (x^{pq}, y^{pq})^T$, i.e.,

$$\psi^{pq}(\mathbf{x}) \equiv \psi_{\text{cs}}\left(\frac{x^{pq} - x}{\Delta x}\right) \cdot \psi_{\text{cs}}\left(\frac{y^{pq} - y}{\Delta x}\right),$$

where $\psi_{\text{cs}}(r)$ is the cubic spline

$$\psi_{\text{cs}}(r) \equiv \begin{cases} \frac{2}{3} - |r|^2 + \frac{1}{2}|r|^3, & |r| \le 1, \\ \frac{1}{6}(2 - |r|)^3, & 1 < |r| \le 2, \\ 0, & |r| > 2. \end{cases}$$

These basis functions form a *partition of unity*, i.e.,

$$\sum_{p,q} \psi^{pq}(\mathbf{x}) = 1.$$

This implies

$$\sum_{p,q} \nabla_{\mathbf{x}} \psi^{pq}(\mathbf{x}) = 0,$$

which is a desirable property when computing gradients. In each layer $i = 1, 2$, we introduce N discrete particles $\mathbf{X}_i^k$, $k = 1, \ldots, N$, with masses m_i^k such that[2]

$$h_i^o(\mathbf{x}^{pq}) \approx \sum_{k=1}^{N} m_i^k \, \psi^{pq}(\mathbf{X}_i^k)$$

at time $t = 0$. More specifically, we approximate the layer-depth h_2 on the grid by

$$h_2^{pq} \equiv \sum_{k=1}^{N} m_2^k \, \psi^{pq}(\mathbf{X}_2^k)$$

and the total layer-depth by

$$h^{pq}(\mathbf{X}) \equiv \sum_{k=1}^{N} \left(m_1^k \, \psi^{pq}(\mathbf{X}_1^k) + m_2^k \psi^{pq}(\mathbf{X}_2^k) \right),$$

where, for later use, we introduced the notation $h^{pq}(\mathbf{X})$ to indicate that h^{pq} depends on all particle positions $\mathbf{X}_i^k$ collected in the vector $\mathbf{X}$.

So far we have essentially followed the standard methodology for deriving *particle-mesh* (PM) methods [10, 4]. The following steps are crucial to the HPM method as introduced in [7] for geophysical fluid dynamics simulations. Even though the layer-depth in rotating fluids often stays relatively smooth, the numerical approximations h_1^{pq} and h^{pq} will develop some nonsmoothness in strongly mixing flows due to the finite number of particles used to resolve the fluid motion; this tends to destabilize PM methods. We suggested in [7] to apply a (discretized) smoothing operator[3]

$$\mathcal{S} = (1 - \alpha^2 \nabla_{\mathbf{x}}^2)^{-p} \tag{2.1}$$

to the layer-depth over the fixed Eulerian grid $\mathbf{x}^{pq}$ with a smoothing length $\alpha = 2\Delta x$ and an exponent $p = 2$. Let us denote the resulting smoothed approximations to h_2^{pq} and h^{pq}, respectively, by $\hat{h}_2^{pq}$ and $\hat{h}^{pq}$. While this smoothing approach has been shown to work very well for compressible flows, it cannot be used to enforce a regularized incompressibility condition (1.3). To see this, note that $\mathcal{S}$ is an invertible operator and, hence, for the (constant) layer-depth approximation,

$$h^{pq}(\mathbf{X}) = \hat{h}^{pq}(\mathbf{X}) = 0.$$

Instead, the following regularization strategy proved successful. We introduce a meta-grid with grid-spacing $\Delta\bar{x} \equiv 2\Delta x$ and grid points denoted by $\bar{\mathbf{x}}^{mn}$. Let $\phi^{mn}(\mathbf{x})$ denote the associated tensor product B-spline centered at $\bar{\mathbf{x}}^{mn} \equiv (\bar{x}^{mn}, \bar{y}^{mn})^T$, i.e.,

$$\phi^{mn}(\mathbf{x}) \equiv \psi_{\rm cs}\left(\frac{\bar{x}^{mn} - x}{\Delta\bar{x}}\right) \cdot \psi_{\rm cs}\left(\frac{\bar{y}^{mn} - y}{\Delta\bar{x}}\right).$$

[2] If the particles $\mathbf{X}_i^k$ are initially placed on a regular grid with equal spacing Δa in the x- and y-direction, then, following (1.4), one can use $m_i^k \equiv h_i^o(\mathbf{X}_i^k)\,(\Delta a/\Delta x)^2$.

[3] In case that the direct numerical implementation of the smoothing operator $\mathcal{S}$ is too expensive, one could replace $\mathcal{S}$ with, e.g., a *Shapiro filter* [23]. In the present paper, we used FFT to implement $\mathcal{S}$.

Then an averaged (coarse-grained) total layer-depth is defined by

$$\bar{h}^{mn}(\mathbf{X}) \equiv \frac{1}{4} \sum_{pq} \phi^{mn}(\mathbf{x}^{pq})\, h^{pq}(\mathbf{X}).$$

The discrete pressure approximation $\bar{p}^{mn}$ is also defined over the coarse grid $\bar{\mathbf{x}}^{mn}$, and the resulting total force acting on particle $\mathbf{X}_i^k$ (excluding the Coriolis contribution) is given by

$$\mathbf{F}_i^k(\mathbf{X}, \bar{\mathbf{p}}) \equiv - \sum_{p,q} \nabla_{\mathbf{X}_i^k} \psi^{pq}(\mathbf{X}_i^k) \times \begin{cases} \left(\sum_{m,n} \phi^{mn}(\mathbf{x}^{pq})\, \bar{p}^{mn} \right), & i = 1, \\ \left(c_0 \hat{h}_2^{pq} + \sum_{m,n} \phi^{mn}(\mathbf{x}^{pq})\, \bar{p}^{mn} \right), & i = 2, \end{cases}$$

where $\bar{\mathbf{p}}$ denotes the vector of pressure variables $\bar{p}^{mn}$. We observe numerically that this coarse graining keeps the pressure gradient sufficiently smooth. This regularization of the pressure field was found necessary for stable computations. Other coarse graining procedures are certainly feasible and will be the subject of further research.

We would like to point out that Holm has revealed in [11] a close relation between velocity smoothing, as fundamental to the α-Euler models [12], and a regularization of the pressure gradient. A closer investigation of our pressure regularization procedure and its relation to the results of [11] is left for future research.

Another important aspect of the HPM method is that the forces are derived from an exact gradient. This implies a number of very desirable conservation properties, such as conservation of circulation, potential vorticity (PV), total mass, and energy [8, 5]. We note that energy conserving variants of PM methods have been considered, for example, in [16] and [14] in the context of plasma physics simulations.

The discrete set of constrained Newtonian equations of motion is now

$$\frac{d}{dt}\mathbf{u}_i^k = f_i^k J \mathbf{u}_i^k + \mathbf{F}_i^k(\mathbf{X}, \bar{\mathbf{p}}), \qquad J \equiv \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \tag{2.2}$$

$$\frac{d}{dt}\mathbf{X}_i^k = \mathbf{u}_i^k, \tag{2.3}$$

$$0 = \bar{h}^{mn}(\mathbf{X}) - H. \tag{2.4}$$

Here f_i^k denotes the value of the Coriolis parameter at particle location $\mathbf{X}_i^k$. In the following, let us first assume that the Coriolis parameter f is constant, i.e., $f = f_0$. Later we will consider the more general case of variable f. Then (2.2)–(2.4) give a constrained Hamiltonian system with the $\bar{p}^{mn}$ variables acting as Lagrange multipliers to enforce the *holonomic constraints* (2.4). The Hamiltonian is

$$\mathcal{H}(\mathbf{X}, \mathbf{v}, \bar{\mathbf{p}}) \equiv \sum_{i=1}^{2} \sum_{k=1}^{N} \frac{1}{2m_i^k} \mathbf{v}_i^k \cdot \mathbf{v}_i^k + \frac{c_0}{2} \sum_{p,q} \hat{h}_2^{pq}(h_2^{pq} - H_2) + \sum_{m,n} (\bar{h}^{mn} - H)\bar{p}^{mn}$$

with conjugate momenta $\mathbf{v}_i^k \equiv m_i^k \mathbf{u}_i^k$. Equations (2.2)–(2.4) are equivalent to

$$\begin{aligned} \frac{d}{dt}\mathbf{v}_i^k &= f_0 J \nabla_{\mathbf{v}_i^k}\mathcal{H} - \nabla_{\mathbf{X}_i^k}\mathcal{H}, \\ \frac{d}{dt}\mathbf{X}_i^k &= \nabla_{\mathbf{v}_i^k}\mathcal{H}, \\ 0 &= \nabla_{\bar{p}^{mn}}\mathcal{H}, \end{aligned}$$

$i = 1, 2$, $k = 1, \ldots, N$. The *symplectic two-form* [2] is given by

$$\omega \equiv \sum_{i,k} \left[d\mathbf{X}_i^k \wedge d\mathbf{v}_i^k + \frac{f_0}{2} d\mathbf{X}_i^k \wedge J^{-1} d\mathbf{X}_i^k \right], \tag{2.5}$$

which is preserved along solutions.

3. Symplectic time-stepping algorithm. Following [13] and [17], we develop a variant of the popular SHAKE/RATTLE algorithm [1, 20, 15] for Hamiltonian systems with holonomic constraints. In particular, the following two steps are performed during each time-step.

Step 1.

$$\mathbf{u}_i^k(t_{n+1/2}) = \mathbf{u}_i^k(t_n) + \frac{\Delta t}{2}\left\{ f_0 J \mathbf{u}_i^k(t_{n+1/2}) + \mathbf{F}_i^k(\mathbf{X}(t_n), \bar{\mathbf{p}}(t_{n+1/2}))\right\}, \tag{3.1}$$

$$\mathbf{X}_i^k(t_{n+1}) = \mathbf{X}_i^k(t_n) + \Delta t \mathbf{u}_i^k(t_{n+1/2}), \tag{3.2}$$

$$0 = \bar{h}^{mn}(\mathbf{X}(t_{n+1})) - H, \tag{3.3}$$

which requires the solution of a nonlinear system in the pressure variable $\bar{\mathbf{p}}(t_{n+1/2})$ to satisfy the holonomic constraint (3.3).

Step 2.

$$\mathbf{u}_i^k(t_{n+1}) = \mathbf{u}_i^k(t_{n+1/2}) + \frac{\Delta t}{2}\left\{ f_0 J \mathbf{u}_i^k(t_{n+1/2}) + \mathbf{F}_i^k(\mathbf{X}(t_{n+1}), \bar{\mathbf{p}}(t_{n+1/2}))\right\}. \tag{3.4}$$

The scheme can be rewritten in terms of the canonical momenta $\mathbf{v}_i^k(t_n)$, and the method conserves the symplectic structure (2.5) from time-step to time-step, i.e., the method is *symplectic* [22]. Backward error analysis [3, 9, 18] implies excellent conservation of energy. It is important that the method used to solve the holonomic constraints (3.3) is iterated to convergence; otherwise, the symplectic property of the algorithm is lost.

If the Coriolis parameter f is not constant, then $f_i^k \equiv f(\mathbf{X}_i^k(t_n))$ is used in (3.1) and $f_i^k \equiv f(\mathbf{X}_i^k(t_{n+1}))$ in (3.4) instead of f_0.

The nonlinear system of equations in the pressure variable $\bar{\mathbf{p}}(t_{n+1/2})$ can be solved by the following quasi-Newton method. Let us denote the iteration index by $l \geq 0$. Then, given some approximation $\bar{\mathbf{p}}^{[l]}$, we can compute the associated approximation to the vector of particle positions $\mathbf{X}^{[l]}(t_{n+1})$ using (3.1)–(3.2) with $\bar{\mathbf{p}}(t_{n+1/2}) = \bar{\mathbf{p}}^{[l]}$. The next pressure approximation

$$\bar{\mathbf{p}}^{[l+1]} \equiv \bar{\mathbf{p}}^{[l]} + \Delta \mathbf{p}^{[l]}$$

is then found by solving

$$A\,\Delta \mathbf{p}^{[l]} = \bar{h}^{mn}(\mathbf{X}^{[l]}(t_{n+1})) - H.$$

The matrix A has entries

$$a^{mn}_{m'n'} \equiv \frac{\Delta t^2}{2} \sum_{k,i,p,q,p',q'} m_i^k \phi^{mn}(\mathbf{x}^{pq}) \left(\nabla \psi^{pq}(\mathbf{X}_i^k) \cdot \nabla \psi^{p'q'}(\mathbf{X}_i^k) \right) \phi^{m'n'}(\mathbf{x}^{p'q'})$$

and is computed only once at the beginning of the simulation with $\mathbf{X}_i^k = \mathbf{X}_i^k(0)$. It is found that the matrix A changes little along the numerical solutions $\mathbf{X}_i^k = \mathbf{X}_i^k(t_n)$. The initial $\bar{\mathbf{p}}^{[0]}$ is found at each time-step from the previous pressure approximation by linear extrapolation, i.e.,

$$\bar{\mathbf{p}}^{[0]} \equiv 2\bar{\mathbf{p}}(t_{n-1/2}) - \bar{\mathbf{p}}(t_{n-3/2}).$$

4. Semi-implicit methods. The step-size of the standard HPM method applied to the two-layer shallow-water equations (1.1) is restricted by the highest frequency of the external gravity waves, which is approximately

$$\omega_{\max} \equiv \sqrt{c_0 H}\sqrt{\frac{g}{g'}}k_{\max},$$

where $k_{\max}$ is the largest computational wave number. This severe step-size restriction motivated the introduction of the constrained HPM method of the previous section. However, let us look back for a moment at the unconstrained shallow-water equations (1.1). The *semi-implicit method*, as pioneered by Robert (see [6]), avoids the step-size restriction of any standard explicit method while also being easy to implement (in particular when combined with a pseudospectral (PS) discretization in space [6]). Let us then briefly review the basic idea behind the semi-implicit method for a one-layer shallow-water model with all advection terms ignored, i.e.,

$$\begin{aligned} \mathbf{u}_t &= f_0 J\mathbf{u} - c\nabla_{\mathbf{x}} h, \qquad c \equiv c_0 \frac{g}{g'}, \\ h_t &= -H\nabla_{\mathbf{x}} \cdot \mathbf{u}. \end{aligned}$$

The (external) *Rossby deformation radius* [21] is equal to

$$\lambda_{\text{ext}} \equiv \frac{\sqrt{cH}}{f_0}.$$

Upon only discretizing in time, the semi-implicit method results in

$$\begin{aligned} \frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1})}{2\Delta t} &= f_0 J\mathbf{u}(t_n) - c\nabla_{\mathbf{x}} \frac{h(t_{n+1}) + h(t_{n-1})}{2}, \\ \frac{h(t_{n+1}) - h(t_{n-1})}{2\Delta t} &= -H\nabla_{\mathbf{x}} \cdot \frac{\mathbf{u}(t_{n+1}) + \mathbf{u}(t_{n-1})}{2}. \end{aligned}$$

Let us define $\hat{h}(t_n) \equiv (h(t_{n+1}) + h(t_{n-1}))/2$. Then, at each time-step, $\hat{h}(t_n)$ is determined by the linear system

$$\left(1 - cH\Delta t^2 \nabla_{\mathbf{x}}^2\right)\hat{h}(t_n) = h(t_{n-1}) - \Delta t H \nabla_{\mathbf{x}} \cdot \mathbf{u}(t_{n-1}) - \Delta t^2 f_0 H \nabla_{\mathbf{x}} \times \mathbf{u}(t_n).$$

This gives

$$\left(1 - cH\Delta t^2 \nabla_{\mathbf{x}}^2\right)\hat{h}(t_n) = h(t_n) + \mathcal{O}(\Delta t^2),$$

which, upon ignoring terms of order $\mathcal{O}(\Delta t^2)$, we will now use as a defining equation for $\hat{h}(t_n)$, i.e.,

$$\hat{h}(t_n) \equiv \mathcal{A}h(t_n), \qquad \mathcal{A} \equiv \left(1 - cH\Delta t^2 \nabla_{\mathbf{x}}^2\right)^{-1}.$$

With this new definition, the semi-implicit method becomes "explicit," i.e.,

$$\begin{aligned}
\frac{\mathbf{u}(t_{n+1}) - \mathbf{u}(t_{n-1})}{2\Delta t} &= f_0 J \mathbf{u}(t_n) - c\nabla_{\mathbf{x}}\mathcal{A}h(t_n),\\
\frac{h(t_{n+1}) - h(t_{n-1})}{2\Delta t} &= -H\nabla_{\mathbf{x}} \cdot \frac{\mathbf{u}(t_{n+1}) + \mathbf{u}(t_{n-1})}{2}.
\end{aligned}$$

This reformulation has the same stability properties as the original semi-implicit method and easily generalizes to the fully nonlinear shallow-water equations.

The important point is that the operator $\mathcal{A}$ becomes equivalent to the smoothing operator (2.1) used in the HPM method for $p = 1$ and $\alpha = \sqrt{cH}\Delta t$. Continuing along this line of thought, we conclude that the HPM method can be made unconditionally stable if used with a smoothing operator

$$\mathcal{S} = \left(1 - cH\Delta t^2 \nabla_{\mathbf{x}}^2\right)^{-1} \tag{4.1}$$

applied to the total layer-depth $h = h_1 + h_2$ in (1.1). However, for Δt severely violating the standard CFL condition [6], we will have $\sqrt{cH}\Delta t \gg \Delta x$, and the time-stepping will lead to excessive smoothing of the total layer-depth. The same argument applies, of course, to the semi-implicit method.

5. Barotropic and baroclinic motion.

Let us introduce the continuous Eulerian velocity approximation

$$\mathbf{u}_1(\mathbf{x}, t) \equiv \frac{\sum_{k=1}^N \mathbf{u}_1^k(t)\psi^{pq}(\mathbf{X}_1^k(t))}{\sum_{k=1}^N \psi^{pq}(\mathbf{X}_1^k(t))}$$

for the first layer, and

$$\mathbf{u}_2(\mathbf{x}, t) \equiv \frac{\sum_{k=1}^N \mathbf{u}_2^k(t)\psi^{pq}(\mathbf{X}_2^k(t))}{\sum_{k=1}^N \psi^{pq}(\mathbf{X}_2^k(t))}$$

for the second layer, respectively. Assuming again that $H_1 = H_2 = 1$, the *barotropic* velocity contribution to the flow is defined by

$$\mathbf{u}(\mathbf{x},t) \equiv \frac{1}{2}\left\{\mathbf{u}_1(\mathbf{x},t) + \mathbf{u}_2(\mathbf{x},t)\right\},$$

which represents synchronized motion in both layers, and the *baroclinic* mode is defined by

$$\Delta\mathbf{u}(\mathbf{x},t) \equiv \frac{1}{2}\left\{\mathbf{u}_1(\mathbf{x},t) - \mathbf{u}_2(\mathbf{x},t)\right\},$$

which represents fluid motions pointing in opposite directions (the *thermal wind*).

If initially $\Delta\mathbf{u} = 0$ and $h_2 = H_2 = 1$, then the *available potential energy* (APE)

$$E_{\text{ap}} \equiv \frac{c_0}{2}\sum_{p,q} \hat{h}_2^{pq}(h_2^{pq} - H_2)$$

is zero and the motion can be reduced to a purely barotropic single layer shallow-water model with a rigid-lid approximation (corresponding to an infinite Rossby deformation radius). On the contrary, $h_2 \neq H_2$ leads to baroclinic motion which is strongly dependent upon its length scale λ relative to the *internal Rossby deformation radius*

$$\lambda_{\text{int}} \equiv \frac{\sqrt{c_0}}{f_0}\sqrt{\frac{H_1 H_2}{H_1 + H_2}} = \sqrt{\frac{c_0}{2f_0^2}}.$$

For length-scales $\lambda \gg \lambda_{\text{int}}$, most of the energy is stored in the layer-depth variation h_2 (i.e., in the APE contribution to $\mathcal{H}$). This energy is eventually transformed into kinetic (barotropic) energy in a process called *baroclinic instability.* In this process the baroclinic modes are reduced to those of length-scale $\lambda \sim \lambda_{\text{int}}$ unless external forcing leads to the activation of large-scale variations in h_2 (such as tropical heating and polar cooling).

Another important concept is that of *geostrophic balance.* By this we mean that the velocities $\mathbf{u}_i$ in each layer stay close to their *geostrophic wind* approximations

$$\mathbf{u}_{\text{gw},1} \equiv f_0^{-1}\nabla_{\mathbf{x}}^{\perp} p, \qquad \mathbf{u}_{\text{gw},2} \equiv f_0^{-1}\nabla_{\mathbf{x}}^{\perp}(p + c_0 h_2)$$

if initialized appropriately. These two definitions imply in particular the balanced thermal wind relation

$$\Delta\mathbf{u}_{\text{thw}} \equiv -\frac{c_0}{2f_0}\nabla_{\mathbf{x}}^{\perp} h_2. \tag{5.1}$$

The associated *baroclinic stream function* $\tau \equiv -c_0/(2f_0)h_2$ represents the vertically averaged temperature anomaly of the fluid.

The *geostrophic approximation* is valid for small *Rossby number* flows, i.e.,

$$Ro \equiv \frac{U}{\lambda f_0} \ll 1,$$

where U and λ are the typical velocity- and length-scales, respectively, for the flow under consideration. For a precise scaling analysis see [21].

To model baroclinic instabilities within the framework of double periodic boundary conditions $\mathbf{x} = (x, y)^T \in \mathcal{R} \equiv [-\pi, +\pi)^2$, we defined a variable Coriolis parameter f by

$$f(y) \equiv f_0 + \beta \sin y.$$

Hence, near $y = 0$, we approximately reproduce a β-plane approximation $f \approx f_0 + \beta y$. See [21] for a detailed explanation of the baroclinic instability.

6. Numerical experiments. We compute the solution starting from a purely baroclinic initial state defined by

$$\mathbf{u}_2 \equiv -\mathbf{u}_1 \equiv \frac{c_0}{2f} \nabla_{\mathbf{x}}^{\perp}(S h_2^o),$$

where $c_0 \equiv 1$, $f \equiv \sqrt{2}(1 + 0.2 \sin y)$,

$$h_2^o(\mathbf{x}) \equiv \frac{1}{1 + 0.08 \exp(-0.85 \|\mathbf{x}\|^2)} + \delta,$$

with the constant δ chosen such that h_2^{pq} has a mean value equal to one. The initial state moves slowly to the left along the x-axis and breaks up into smaller (barotropic and baroclinic) vortices. The internal deformation radius is $\lambda_{\text{int}} = 0.5$.

The spatial grid resolution for the rigid-lid HPM method is $\Delta x = 2\pi/128 \approx 0.0491$ with $N = 262144$ particles per layer, i.e., $\Delta a = \Delta x/4$. The smoothing length in (2.1) is $\alpha = 2\Delta x \approx 0.0982$ and the operator $\mathcal{S}$ is implemented using an FFT. We also implemented a PS method for the standard Eulerian formulation of the compressible two-layer shallow-water equations with $\Delta x = 2\pi/256 \approx 0.0245$ and a semi-implicit discretization in time (see [6]). We stress that no hyperdiffusion was applied. The external deformation radius for the unconstrained shallow-water model is $\lambda_{\text{ext}} = 20$, i.e., $g' = g/400$ and $c = 400$.

Both methods were implemented using MATLAB, and `mex`-files were used for the PM computations within the HPM method. Note that $\sqrt{g'/g} = 20$ implies that a standard HPM discretization of the unfiltered equation (1.1) would require a step-size about 30 times smaller than the rigid-lid HPM method. This severe step-size restriction does not apply to the semi-implicit PS method. However, it was found that the largest possible step-size for the rigid-lid HPM method is $\Delta t = 0.5$, while the semi-implicit PS method requires $\Delta t \leq 0.07$ to be stable for the given initial data and $t \in [0, 150]$.

Figure 6.1 shows the time evolution of the baroclinic and barotropic vorticity fields over a time interval $[0, 150]$ using a step-size of $\Delta t = 0.1$. The corresponding results from the semi-implicit PS method with step-size $\Delta t = 0.01$ and initial $h_1^o \equiv 2 - h_2^o$ can be found in Figure 6.2. Note that this step-size leads to a computational smoothing length of $\alpha \approx 0.2828 \gg \Delta x \approx 0.0245$ in (4.1). We recorded the CPU-time for both simulations and obtained about 36000 time-units for the HPM method and 48000 time-units for the semi-implicit PS method. For output purposes, the smoothing operator (2.1) was applied to the gridded vorticity fields to average out fine-scale vorticity filaments. The vorticity fields are identical up to some small-scale differences over the whole time interval $[0, 150]$.

We also prepared a few videos using the GIF format. One can access these by clicking one of the following four options:

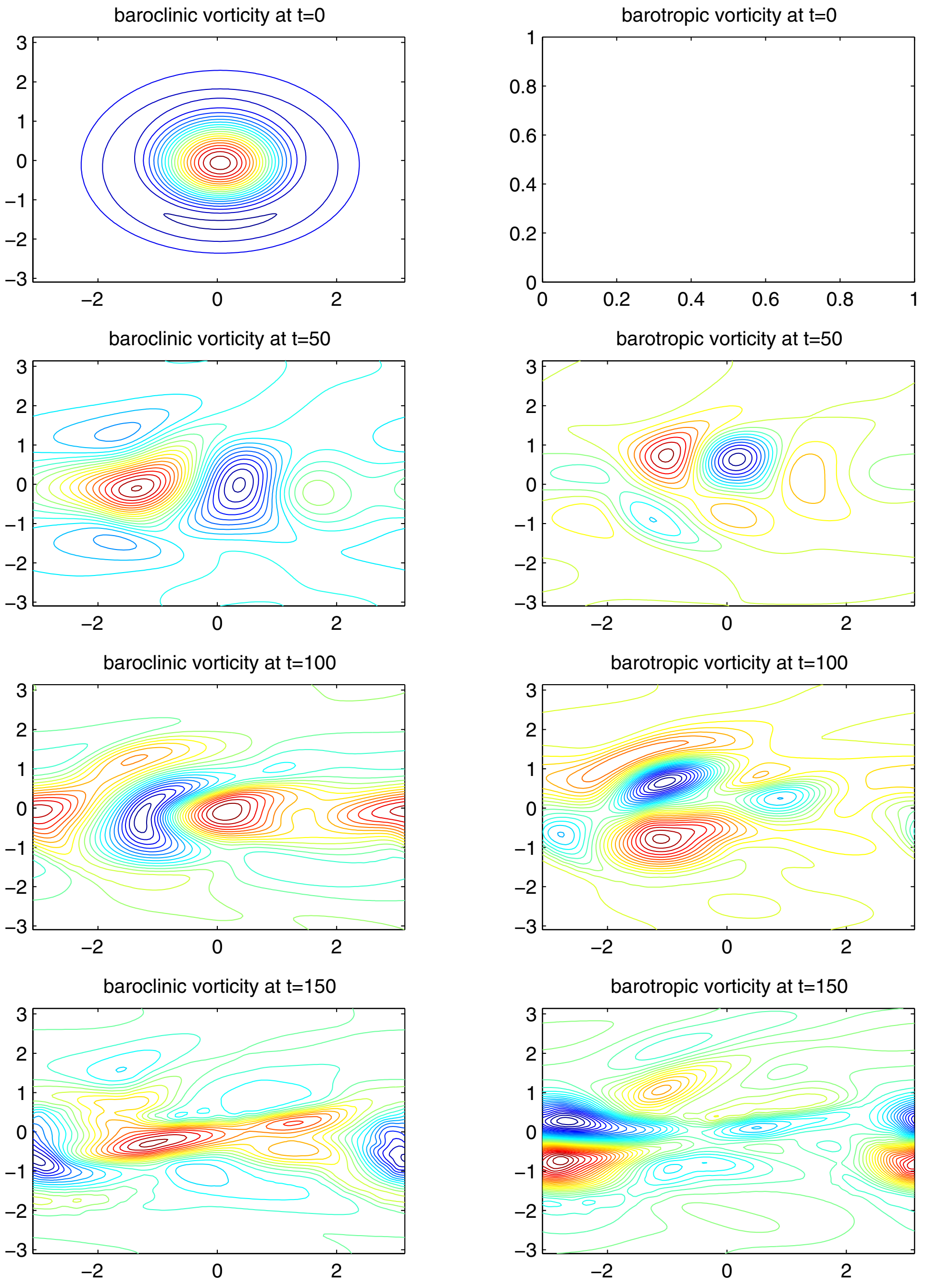


Figure 6.1. *HPM simulation for shallow-water model with rigid lid. Top to bottom: time evolution of vorticity. Left: baroclinic vorticity. Right: barotropic vorticity.*

(i) particle motion in top layer,
(ii) particle motion in bottom layer,
(iii) baroclinic vorticity field,
(iv) barotropic vorticity field.

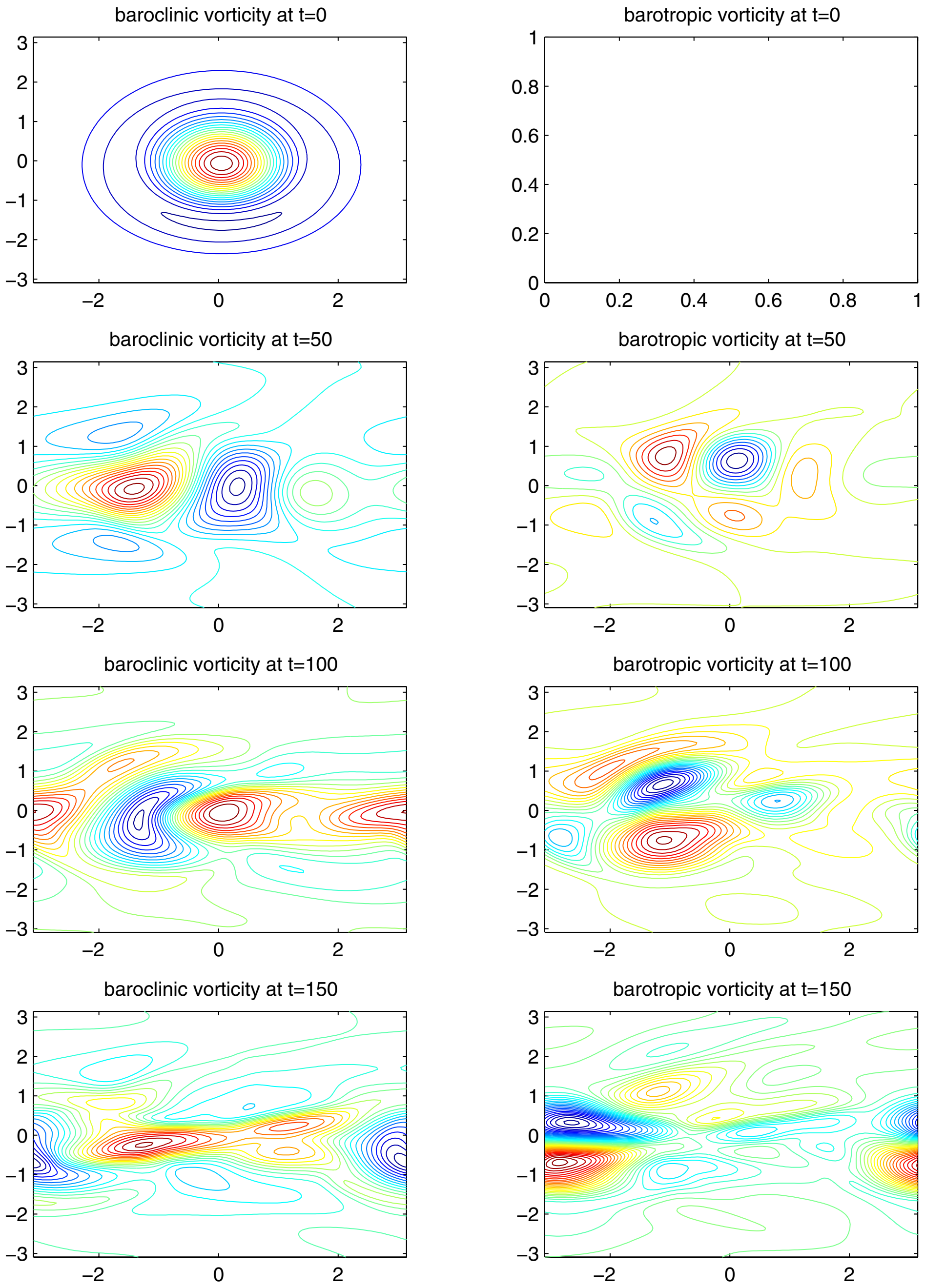


Figure 6.2. *PS simulation for shallow-water model without rigid-lid. Top to bottom: time evolution of vorticity; left: baroclinic vorticity; right: barotropic vorticity.*

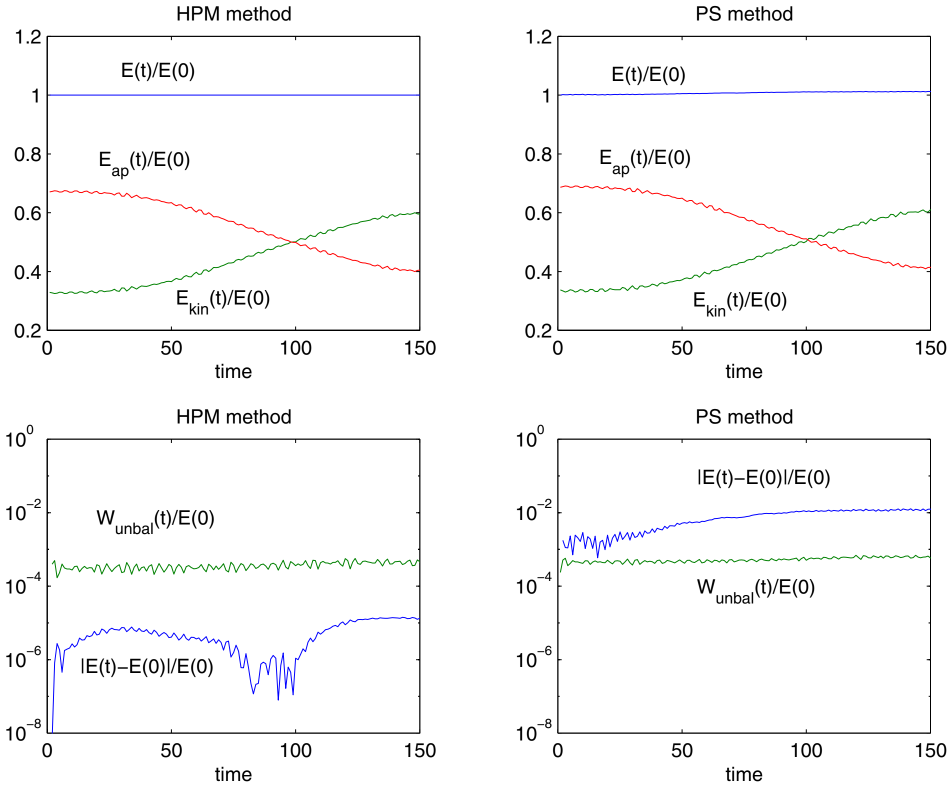


Figure 6.3. *Diagnostic results.*

A few diagnostic results for the rigid-lid and unconstrained simulations can be found in Figure 6.3. More specifically, let $E(t_n)$ denote the total energy of the PM model,

$$E_{\rm kin}(t_n) \equiv \sum_{i=1}^{2}\sum_{k=1}^{N} \frac{1}{2m_i^k}\mathbf{v}_i^k(t_n)\cdot\mathbf{v}_i^k(t_n)$$

its kinetic energy (KE), and

$$E_{\rm ap}(t_n) \equiv \frac{c_0}{2}\sum_{p,q} \hat{h}_2^{pq}(t_n)[h_2^{pq}(t_n) - H_2]$$

its APE. For the incompressible rigid-lid model, we have $E(t_n) = E_{\rm kin}(t_n) + E_{\rm ap}(t_n)$. Up to a small potential energy contribution from the total layer-depth, this is essentially also true for the compressible two-layer model. We plot in Figure 6.3 the scaled quantities $E(t_n)/E(t_0)$, $E_{\rm kin}(t_n)/E(t_0)$, and $E_{\rm ap}(t_n)/E(t_0)$ with $t_0 = 0$. Furthermore, we also monitor the norm of the unbalanced baroclinic velocity contributions

$$W_{\rm unbal}(t_n) \equiv \frac{1}{2}\|\Delta\mathbf{u}(t_n) - \Delta\mathbf{u}_{\rm thw}(t_n)\|_2^2,$$

with $\Delta\mathbf{u}_{\rm thw}$ defined by (5.1). Here all velocities are first approximated over the grid $\mathbf{x}^{pq}$, and then $\|.\|_2$ is to be understood as the discrete l_2-norm. The scaled variable $W_{\rm unbal}(t_n)/E(t_0)$

and the numerically induced errors in total energy can be found in Figure 6.3. The quasi-conservation of balanced motion for both methods, as manifested by the very small ratio $W_{\rm unbal}(t_n)/E(t_0)$, is particularly striking. The Rossby number for the simulation was $Ro \approx 0.1 - 0.2$. We also observe that the particle method conserves total energy much better than the semi-implicit PS method.

We would like to point out that the given initial purely baroclinic state is persistent in the absence of the β-plane effect. Hence the break-up of the initial state into baroclinic and barotropic motions is triggered by $\beta \neq 0$.

7. Conclusions. Three dominant themes within geophysical fluid dynamics are (i) conservation, (ii) model reduction, and (iii) multiscales. A simple model system that combines all three of these aspects is provided by the two-layer shallow-water equations. These equations are Hamiltonian, satisfy conservation laws of PV and circulation, and can be simplified by filtering out surface gravity waves via the rigid-lid approximation. Geostrophic balance is of utmost importance for the long-time solution behavior in a small Rossby number regime. In the present paper, we have demonstrated how these ideas and concepts can be filtered through to the level of numerical methods. The proposed discrete PM method is Hamiltonian and conserves circulation/PV along the lines of [8, 5]. Furthermore, symplectic time-stepping guarantees maintenance of geostrophic balance as an adiabatic invariant [19]. Finally, the rigid-lid approximation is implemented as a holonomic constraint which allows significant increases in the attainable time-steps. A coarse graining procedure has been implemented to keep the numerical pressure gradient sufficiently smooth. This technique appears to be related to regularization techniques discussed in [11].

We hope that the presented PM method can serve as a role model for further developments on more realistic model systems such as the primitive equations (see [21]).

REFERENCES

[1] H. Andersen, *Rattle: A ‘velocity’ version of the shake algorithm for molecular dynamics calculations*, J. Comput. Phys., 52 (1983), pp. 24–34.

[2] V. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer, New York, 1988.

[3] G. Benetin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys., 74 (1994), pp. 1117–1143.

[4] C. Birdsall and A. Langdon, *Plasma Physics via Computer Simulations*, McGraw-Hill, New York, 1981.

[5] T. Bridges, P. Hydon, and S. Reich, *Vorticity and Symplecticity in Lagrangian Fluid Dynamics*, Tech. report, University of Surrey, Guildford, 2002.

[6] D. Duran, *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*, Springer, New York, 1999.

[7] J. Frank, G. Gottwald, and S. Reich, *The Hamiltonian particle-mesh method*, in Meshfree Methods for Partial Differential Equations, Lecture Notes in Comput. Sci. Engrg. 26, M. Griebel and M. Schweitzer, eds., Springer, Heidelberg, 2002, pp. 131–142.

[8] J. Frank and S. Reich, *Conservation properties of smoothed particle hydrodynamics applied to the shallow-water equations*, BIT, 43 (2003), pp. 40–54.

[9] E. Hairer and C. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math., 76 (1997), pp. 441–462.

[10] R. Hockney and J. Eastwood, *Computer Simulations Using Particles*, Institute of Physics, Philadelphia, 1988.

[11] D. Holm, *Fluctuation effects on 3D Lagrangian mean and Eulerian mean fluid motion*, Phys. D, 133 (1999), pp. 215–269.

[12] D. Holm, J. Marsden, and T. Ratiu, *Euler-Poincaré models of ideal fluids with nonlinear dispersion*, Phys. Rev. Lett., 80 (1998), pp. 4173–4177.

[13] L. Jay, *Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems*, SIAM J. Numer. Anal., 33 (1996), pp. 368–387.

[14] A. Langdon, *Energy-conserving plasma simulation algorithms*, J. Comput. Phys., 12 (1973), pp. 247–268.

[15] B. Leimkuhler and R. Skeel, *Symplectic numerical integrators in constrained Hamiltonian systems*, J. Comput. Phys., 112 (1994), pp. 117–125.

[16] H. Lewis, *Energy-conserving numerical approximations for Vlasov plasmas*, J. Comput. Phys., 6 (1970), pp. 136–141.

[17] S. Reich, *Symplectic integration of constrained Hamiltonian systems by composition methods*, SIAM J. Numer. Anal., 33 (1996), pp. 475–491.

[18] S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal., 36 (1999), pp. 1549–1570.

[19] S. Reich, *Conservation of adiabatic invariants under symplectic discretization*, Appl. Numer. Math., 29 (1999), pp. 45–55.

[20] J. Ryckaert, G. Ciccotti, and H. Berendsen, *Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes*, J. Comput. Phys., 23 (1977), pp. 327–341.

[21] R. Salmon, *Geophysical Fluid Dynamics*, Oxford University Press, Oxford, UK, 1999.

[22] J. Sanz–Serna and M. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, London, 1994.

[23] R. Shapiro, *Smoothing, filtering and boundary effects*, Rev. Geophys. Space Phys., 8 (1970), pp. 359–387.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 2, pp. 85–116

Freezing Solutions of Equivariant Evolution Equations[*]

W.-J. Beyn[†] and V. Thümmler[†]

Abstract. In this paper we develop numerical methods for integrating general evolution equations $u_t = F(u)$, $u(0) = u_0$, where F is defined on a dense subspace of some Banach space (generally infinite-dimensional) and is equivariant with respect to the action of a finite-dimensional (not necessarily compact) Lie group. Such equations typically arise from autonomous PDEs on unbounded domains that are invariant under the action of the Euclidean group or one of its subgroups. In our approach we write the solution $u(t)$ as a composition of the action of a time-dependent group element with a "frozen solution" in the given Banach space. We keep the frozen solution as constant as possible by introducing a set of algebraic constraints (phase conditions), the number of which is given by the dimension of the Lie group. The resulting PDAE (partial differential algebraic equation) is then solved by combining classical numerical methods, such as restriction to a bounded domain with asymptotic boundary conditions, half-explicit Euler methods in time, and finite differences in space. We provide applications to reaction-diffusion systems that have traveling wave or spiral solutions in one and two space dimensions.

Key words. general evolution equations, equivariance, Lie groups, partial differential algebraic equations, unbounded domains, asymptotic boundary conditions

AMS subject classifications. 65M99, 35K57

DOI. 10.1137/030600515

1. Introduction.

We consider the numerical solution of general evolution equations

$$u_t = F(u), \quad u(0) = u_0, \tag{1.1}$$

that are equivariant with respect to the action of a finite-dimensional, not necessarily compact Lie group. Equation (1.1) is considered on a Banach space X, where the mapping F has a dense domain. Equivariance means that we have a finite-dimensional Lie group G, which acts on X via a representation $a : G \mapsto GL(X)$ such that F is equivariant in the sense $F(a(\gamma)v) = a(\gamma)F(v)$ for all $\gamma \in G$ and for all v in the domain of F.

The main applications we have in mind are reaction-diffusion systems on unbounded domains $\Omega \subset \mathbb{R}^d$ such as the semilinear system

$$u_t = \Delta u + f(u), \quad x \in \Omega, \quad u(0) = u_0, \tag{1.2}$$

where $u(x,t) \in \mathbb{R}^m$, and $f : \mathbb{R}^m \mapsto \mathbb{R}^m$ is sufficiently smooth. If the domain Ω is invariant with respect to the action of a Lie group $G \subset GL(\mathbb{R}^d)$, then this induces an equivariance

[*] Received by the editors July 21, 2003; accepted for publication (in revised form) by M. Golubitsky November 26, 2003; published electronically May 19, 2004. This work was supported by the DFG research group "Spectral analysis, asymptotic distributions and stochastic dynamics," Bielefeld University.

http://www.siam.org/journals/siads/3-2/60051.html

[†] Fakultät für Mathematik, Universität Bielefeld, Postfach 100131, D-33501 Bielefeld, Germany (beyn@mathematik.uni-bielefeld.de, thuemmle@mathematik.uni-bielefeld.de).

of (1.2) via the action

$$[a(\gamma)v](x) = v(\gamma x), \quad x \in \Omega,$$

where v is in some suitable function space. In the case of $\Omega = \mathbb{R}^d$, equivariance holds with respect to the Euclidean group $G = SE(d)$. Further symmetries may be induced by special equivariance properties of the linear and nonlinear parts in (1.2).

Up to now there is a well-developed bifurcation theory for equivariant dynamical systems that covers the infinite-dimensional case of PDEs and certain aspects of noncompact Lie groups; see the monographs [12], [6]. In particular, we refer to [13], [23], and the remarkable series of papers [10], [28], [29], [11]. One of the underlying ideas in the latter papers is to transform the flow of (1.1) into a so-called skew product form. One part is orthogonal to the group orbit (of the initial value) and the other part acts within the group orbit and depends upon the position in the orthogonal direction; compare [10]. Combining this decomposition with center manifold reductions (see [28], [29]) leads to a powerful tool for studying equivariant bifurcations in PDEs. In this way, various bifurcations of spiral waves, observed and interpreted in [1], [3], could be put into a mathematically rigorous framework.

In this paper we propose a numerical method for solving the initial value problem that makes use of the equivariance by extending the system (1.1) rather than reducing it as in bifurcation analysis. More precisely, we write the solution $u(t)$ of (1.1) as

$$u(t) = a(\gamma(t))v(t), \tag{1.3}$$

where $\gamma(t) \in G$ and $v(t) \in X$ are to be determined. The extra degrees of freedom $\gamma(t)$ are compensated for by phase conditions

$$\psi(v, \gamma) = 0, \tag{1.4}$$

the number of which is given by the dimension of the Lie group. The resulting system for $(v(t), \gamma(t))$ (see (2.18)) is an abstract differential algebraic equation, which will be set up and analyzed in some detail in section 2. The choice of phase condition is crucial for our approach since it determines the parametrization of the v-orbits.

In section 2.3 we discuss several choices for the function ψ in (1.4) that are based on minimization or orthogonality principles. In particular, near relative equilibria of (1.1) (i.e., solutions of the form $u(t) = a(\gamma(t))v$) the phase condition should force the v-part of the solution to become stationary. For this reason we will sometimes call $v(t)$ the *frozen solution* and the transformed system the *frozen system*.

Applications to parabolic systems (1.2) in one and two space dimensions will be discussed in sections 2 and 3. The frozen system in this case turns out to be a PDAE (partial differential algebraic equation), which will be solved in a straightforward manner by a half-explicit Euler method. For numerical computations one has to restrict the infinite to a finite domain and use appropriate boundary conditions. After this truncation the original and the frozen systems are no longer equivalent. For example, when a traveling wave (or a drifting spiral) reaches a finite boundary in the given system, it will usually die out, while in the frozen system it is expected to become stationary.

In section 3 we will discuss several two-dimensional systems from the literature (e.g., Barkley's spiral system [1], [3], the $(\lambda-\omega)$-system [18], and the quintic Ginzburg–Landau equation [7], [8]) that show rigidly rotating spiral waves. Freezing such waves can be delicate because it depends on the precise choice of phase condition (with or without weighted L^2-norms), the type of numerical discretization (rectangular or polar grid), and on the right choice of the underlying group. Note that a related approach to ours was developed in [5] with the intention of using the side constraint in order to fix the tip of a spiral wave. Also in [2] Barkley mentions the use of a pinning condition or phase condition in order to compute the spiral wave from a time-independent boundary value problem.

While there have been quite a few numerical bifurcation methods that employ equivariance with respect to compact and mostly discrete groups (see [14, Ch. 8] for a recent survey) it seems that equivariance with respect to general Lie groups has not been systematically used for solving equivariant systems numerically. We expect that, apart from the evolution system (1.1), our general approach will also be useful for the numerical bifurcation analysis of relative equilibria and relative periodic orbits.

After preparation of this manuscript we learned of the related work of Rowley et al. [24], which builds on previous work by Rowley and Marsden [25]. In [25] the idea of splitting the solutions in the form (2.16) and adding a minimization condition like (2.26) appears in the context of Karhunen–Loève expansion for systems with symmetry. Equation (2.18) is then called the reconstruction equation. In [24] the authors generalize this approach to dynamical systems with self-similar symmetries, which, in addition to the equivariance used in this paper, allow rescalings of the time variable. Applications to traveling waves in one space dimension (Kuramoto–Shivashinsky, Burgers) are presented in [25], [24].

2. The general approach.

2.1. Equivariant evolution equations.

In this section we set up the technique of decomposing the solutions of the evolution equation (1.1) in an abstract setting. Simultaneously, we treat two important examples (parabolic systems on the line and in the plane) in a formal way with the details of a proper functional analytic setting given in the subsequent sections. We assume that $(X, ||\cdot||)$ is a Banach space and Y is a dense subspace on which the operator F from (1.1) is defined, i.e.,

$$F : \begin{array}{c} Y \subset X \to X, \\ \qquad u \mapsto F(u), \end{array} \qquad \overline{Y} = X. \tag{2.1}$$

Example 2.1. Consider the parabolic system

$$u_t = Au_{xx} + f(u, u_x) =: F(u), \quad -\infty < x < \infty, \tag{2.2}$$

where $u(x,t) \in \mathbb{R}^m$, A is a positive definite $m \times m$ matrix and $f : \mathbb{R}^{2m} \to \mathbb{R}^m$ is assumed to be sufficiently smooth. If $f(0) = 0$, and if f and its first derivative are globally bounded, then (2.1) holds for the choice

$$Y = H^2(\mathbb{R}, \mathbb{R}^m), \quad X = L^2(\mathbb{R}, \mathbb{R}^m). \tag{2.3}$$

Clearly, this excludes solutions that do not decay at $\pm\infty$. A more general setting will be discussed in sections 2.4 and 3.

In this example and in what follows we use $H^k(\mathbb{R},\mathbb{R}^m)$, $k \geq 1$, to denote the standard Sobolev space of functions that have generalized derivatives in L^2 up to order $k \geq 1$.

As a second example we mention the semilinear equation

$$u_t = A\Delta u + f(u) =: F(u), \quad x \in \mathbb{R}^2, \tag{2.4}$$

where $u(x,t)$ and A are as above and $f : \mathbb{R}^m \to \mathbb{R}^m$ satisfies appropriate smoothness and boundedness assumptions (see section 4).

We further assume that a finite-dimensional (not necessarily compact) Lie group $(G,\circ)$ is given that acts on X via a representation in $GL(X)$; that is, we have a homomorphism (cf. [6, Ch. 4.3.1])

$$a : \begin{array}{l} G \to GL(X) \\ \gamma \mapsto a(\gamma) \end{array} \tag{2.5}$$

satisfying

$$a(\mathbb{1}) = I, \quad a(\gamma_1 \circ \gamma_2) = a(\gamma_1)a(\gamma_2). \tag{2.6}$$

Here $\mathbb{1}$ and I denote the unit elements in G and $GL(X)$, respectively.

Our main assumption is that the mapping F from (2.1) is *equivariant under the action of* G in the following sense.

Hypothesis 2.2. *For all* $\gamma \in G$*,*

$$a(\gamma)(Y) \subseteq Y, \tag{2.7}$$

$$F(a(\gamma)u) = a(\gamma)F(u) \quad \forall u \in Y. \tag{2.8}$$

In the last equation we restrict the equivariance condition to the dense subspace Y. Equation (2.7) is included to ensure that $a(\gamma)$ maps the domain of F into itself. Note that (2.7) implies $a(\gamma)(Y) = Y$ since for every $\gamma \in G$,

$$Y = a(\gamma)a(\gamma^{-1})(Y) \subseteq a(\gamma)(Y) \subseteq Y.$$

Example 2.3. Consider Example 2.1 with the additive group $(G,\circ) = (\mathbb{R},+)$ and the action $a(\gamma)$ defined by the shift

$$[a(\gamma)u]\,(x) := u(x-\gamma), \quad x \in \mathbb{R}, \gamma \in G. \tag{2.9}$$

With the spaces from (2.3), (2.7) and (2.8) are satisfied.

In case (2.4) we take the two-dimensional Euclidean group (see [6], [28], [29], [10])

$$G = SE(2) = S^1 \ltimes \mathbb{R}^2,$$

where the semidirect product $S^1 \ltimes \mathbb{R}^2$ is defined topologically by the direct product

$$\gamma = (\theta, b) \in S^1 \times \mathbb{R}^2,$$

and the group operation is given by

$$\gamma_1 \circ \gamma_2 = (\theta_1, b_1) \circ (\theta_2, b_2) = (\theta_1 + \theta_2, b_1 + \varrho_{\theta_1} b_2). \tag{2.10}$$

Here the unit element is $\mathbb{1} = (0, 0)$ and we write rotations in $\mathbb{R}^2$ as

$$\varrho_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{2.11}$$

The action on functions is given by (see [6], [29], [10])

$$[a(\gamma)u](x) := u(\varrho_{-\theta}(x - b)), \quad x \in \mathbb{R}^2. \tag{2.12}$$

One easily verifies the property (2.6) and the equivariance condition (2.8) with the help of the Euclidean equivariance of the Laplacian

$$\Delta_x [u(\varrho_\theta x)] = \Delta u(\varrho_\theta x), \quad \Delta_x [u(x - b)] = \Delta u(x - b).$$

2.2. Separating the group motion. A well-known problem in the infinite-dimensional setting is differentiability of the group action (see [28], [29], [10] for a detailed discussion). We can neither expect the mapping a to be differentiable from G into $GL(X)$ nor assume that the mapping $\gamma \mapsto a(\gamma)u$ is differentiable for any fixed $u \in X$. Our assumption is as follows.

Hypothesis 2.4. *For any $v \in X$, the mapping*

$$a(\cdot)v : \begin{array}{c} G \to X \\ \gamma \mapsto a(\gamma)v \end{array} \tag{2.13}$$

is continuous, and for any $v \in Y$ it is continuously differentiable with derivative

$$a_\gamma(\gamma)v : \begin{array}{c} T_\gamma G \to X \\ \lambda \mapsto [a_\gamma(\gamma)v]\,\lambda. \end{array} \tag{2.14}$$

We use $\mathcal{A}_\gamma = T_\gamma G$ to denote the tangent space of G at γ. Note that $\mathcal{A} := \mathcal{A}_{\mathbb{1}}$ is the Lie algebra associated with G, which has the same dimension as G. A general principle of constructing spaces that satisfy Hypothesis 2.4 will be discussed in section 2.4.

For Example 2.3, continuity is satisfied for $v \in L^2(\mathbb{R}, \mathbb{R}^m)$, and continuous differentiability holds for functions $v \in H^1(\mathbb{R}, \mathbb{R}^m) \supset Y = H^2(\mathbb{R}, \mathbb{R}^m)$ with

$$[a_\gamma(\gamma)v]\,\lambda = -v_x(\cdot - \gamma)\lambda. \tag{2.15}$$

Consider a solution $u(t)$ of (1.1) and a function $\gamma \in C^1(\mathbb{R}, G)$, $\gamma(0) = \mathbb{1}$, and define $v(t)$ via (see Figure 2.1)

$$u(t) = a(\gamma(t))v(t). \tag{2.16}$$

Then by differentiating formally and using the equivariance condition, we obtain

$$u_t = [a_\gamma(\gamma)v]\,\gamma_t + a(\gamma)v_t = F(u) = F(a(\gamma)v) = a(\gamma)F(v). \tag{2.17}$$

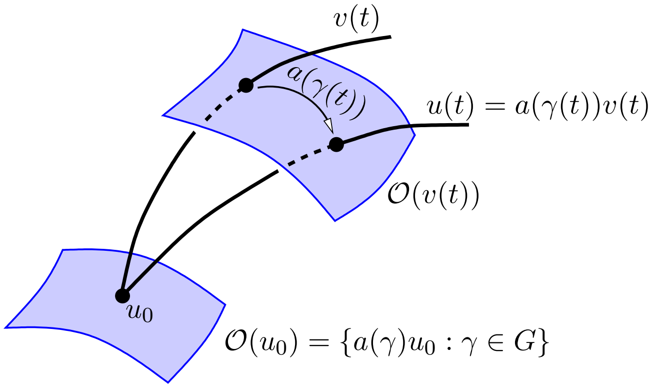

$v(t)$
$a(\gamma(t))$
$u(t) = a(\gamma(t))v(t)$
$\mathcal{O}(v(t))$
u_0
$\mathcal{O}(u_0) = \{a(\gamma)u_0 : \gamma \in G\}$

Figure 2.1. *Splitting off the group dynamics by extension.*

Applying $a(\gamma^{-1}) = a(\gamma)^{-1}$ to both sides, we end up with the equation for the "frozen solution" $v(t)$:

$$v_t = F(v) - a(\gamma^{-1})a_\gamma(\gamma)v\gamma_t, \quad v(0) = u_0. \tag{2.18}$$

In order to make the equivalence of (1.1) and (2.18) rigorous we use a working definition for solutions, which shares at least some properties of strong solutions for evolution equations with C^0 semigroups (see [17], [28], [29], [10] for the standard solution concept in fractional order spaces when $F = A + f$ with a sectorial operator A).

Definition 2.5. *A function $u \in C([0,T), X) \cap C^1((0,T), X)$ is called a solution of the initial value problem (1.1) on $[0,T)$ if $u(0) = u_0$, $u(t) \in Y$ for $0 < t < T$ and if the differential equation holds in the open interval $(0,T)$.*

With this notion we obtain the following theorem.

Theorem 2.6. *Suppose that Hypotheses 2.2 and 2.4 hold and let $\gamma \in C^1([0,T), G)$ satisfy $\gamma(0) = \mathbb{1}$. Then u is a solution of (1.1) if and only if v, given by (2.16), is a solution of (2.18).*

Proof. We may write (2.16) equivalently as $v(t) = a(\gamma(t)^{-1})u(t)$, where $\gamma^{-1} \in C^1([0,T), G)$ and $\gamma^{-1}(0) = \mathbb{1}$. Therefore, it is sufficient to show that $u(t)$ inherits the smoothness from $v(t)$ and that the first equality in (2.17) holds. As in semigroup theory, one concludes from Hypothesis 2.4 and the uniform boundedness principle that the operator norms $|a(\gamma)|$ are uniformly bounded when γ varies in a compact set. This implies continuity of the map $(\gamma, u) \mapsto a(\gamma)u$ on $G \times X$ and thus continuity of $u(t)$ for $t \in [0,T)$. Differentiability in $(0,T)$ and the desired formula follow from Hypothesis 2.4 by a careful look at the standard proof of the chain and product rule:

$$\begin{aligned} u(t+h) - u(t) &= a(\gamma(t+h))(v(t+h) - v(t) - v_t(t)h) + a(\gamma(t))v_t(t)h \\ &\quad + (a(\gamma(t+h)) - a(\gamma(t)))v_t(t)h + (a(\gamma(t+h)) - a(\gamma(t)))v(t) \\ &= a(\gamma(t))v_t(t)h + [a_\gamma(\gamma(t))v(t)]\,\gamma_t(t)h + o(h). \qquad ■ \end{aligned}$$

In (2.18) the path $\gamma(t)$ in the group is still arbitrary. We fix these degrees of freedom by so-called "phase conditions," the number of which equals the dimension of the group. We assume that we are given a map

$$\psi : \begin{array}{c} X \times G \to \mathcal{A}^* \\ (u, \gamma) \mapsto \psi(u, \gamma) \end{array} \tag{2.19}$$

that satisfies the consistency relation $\psi(u_0, \mathbb{1}) = 0$. In (2.19) we use $\mathcal{A}^*$ to denote the dual of the Lie algebra $\mathcal{A}$. Further, by introducing the variable $\lambda = \gamma_t$ we finally arrive at the following PDAE for the time-dependent variables $\gamma(t) \in G$, $\lambda(t) \in T_{\gamma(t)}G$, $v(t) \in Y$:

$$v_t = F(v) - a(\gamma^{-1})a_\gamma(\gamma)v\lambda, \qquad v(0) = u_0, \tag{2.20}$$

$$\gamma_t = \lambda, \qquad \gamma(0) = \mathbb{1}, \tag{2.21}$$

$$0 = \psi(v, \gamma). \tag{2.22}$$

In general (2.20) is a PDE, (2.21) an ODE system on a manifold, and (2.22) is an algebraic constraint. In our example (2.2), (2.3) the PDAE reads

$$v_t = Av_{xx} + f(v, v_x) + v_x\lambda, \qquad v(0) = u_0, \tag{2.23}$$

$$\gamma_t = \lambda, \qquad \gamma(0) = 0, \tag{2.24}$$

$$0 = \psi(v, \gamma). \tag{2.25}$$

Remark 2.7. Decomposing the solution as in (2.16) is also the underlying idea in the center manifold reduction in [28], [29], [10] as well as in the slice theorem, see [6, Ch. 6]. However, rather than using it to derive a reduced system which contains global terms from elimination, we set up an extended system that keeps most of the structure of the original problem. This will be better suited for numerical methods.

We extend Definition 2.5 by saying that the tuple (v, γ, λ) is a solution of the PDAE on $[0, T)$ if $v \in C([0, T), X) \cap C^1((0, T), X)$ and $(\gamma, \lambda) \in C^1([0, T), TG)$ such that $v(t) \in Y$ for $0 < t < T$ and such that the differential equation in (2.20) holds in $(0, T)$ and the equations in (2.21), (2.22) hold in $[0, T)$. Note that the tangent bundle TG consists of pairs (γ, λ) with $\lambda \in T_\gamma G$.

For the phase condition we use the following.

Hypothesis 2.8. *$\psi \in C^1(X \times G, \mathcal{A}^*)$, $\psi(u_0, \mathbb{1}) = 0$ and the linear map*

$$\psi_\gamma(u_0, \mathbb{1}) - \psi_v(u_0, \mathbb{1})a_\gamma(\mathbb{1})u_0 : \mathcal{A} \mapsto \mathcal{A}^*$$

is nonsingular.

The first part of the following theorem is an immediate consequence of Theorem 2.6.

Theorem 2.9. *If (v, γ, λ) is a solution of (2.20)–(2.22) on $[0, T)$, then $u(t) = a(\gamma(t))v(t)$ solves (1.1). Conversely, assume that $u(t)$ solves (1.1) on some interval $[0, T)$ and that ψ satisfies Hypothesis 2.8. Then there exists an interval $[0, \tau) \subset [0, T)$ and a function $\gamma \in C^1([0, \tau), G)$ such that $(v = a(\gamma^{-1})u, \gamma, \lambda = \gamma_t)$ is a solution of (2.20)–(2.22) on $[0, \tau)$.*

Proof. For the second part apply the implicit function theorem to the equation

$$\varphi(\gamma, t) := \psi(a(\gamma^{-1})u(t), \gamma) = 0.$$

Note that $\varphi(\mathbb{1}, 0) = 0$ and that $\varphi_\gamma(\mathbb{1}, 0) = \psi_\gamma(u_0, \mathbb{1}) - \psi_v(u_0, \mathbb{1})a_\gamma(\mathbb{1})u_0$ is invertible by Hypothesis 2.8. ■

2.3. Phase conditions. Let us assume that we have an inner product $\langle \cdot, \cdot \rangle$ on X that is continuous with respect to the given norm $|| \cdot ||$, i.e., $|u| = \sqrt{\langle u, u \rangle} \leq C||u||$. For (2.3) in Example 2.1, the two norms are identical and X is a Hilbert space, but we do not assume this is general. In later applications we will use weighted and locally uniform norms for which the two norms differ.

One way to set up a phase condition is to minimize the distance of the frozen solution v from the group orbit of the starting value

$$\mathcal{O}(u_0) = \{a(\gamma)u_0 : \gamma \in G\},$$

i.e., minimize $e_1(\gamma) = |a(\gamma)u_0 - v|^2$. If we require u_0 to be the point on the orbit that is closest to v, we obtain from Hypothesis 2.4 the necessary condition

$$\psi_1(v, \gamma)\mu := \langle a_\gamma(\mathbb{1})u_0\mu, u_0 - v \rangle = 0 \quad \forall \mu \in \mathcal{A}. \tag{2.26}$$

Similarly, we may require that v is the point of minimal distance from u_0 on $\mathcal{O}(v)$; i.e., we minimize $e_2(\gamma) = |u_0 - a(\gamma)v|^2$. This leads to (see Figure 2.2 for an illustration)

$$\psi_2(v, \gamma)\mu := \langle a_\gamma(\mathbb{1})v\mu, u_0 - v \rangle = 0 \quad \forall \mu \in \mathcal{A}. \tag{2.27}$$

Proposition 2.10. *If the isotropy subgroup (or stabilizer) of u_0, given by*

$$\mathit{Stab}(u_0) = \{\gamma \in G : a(\gamma)u_0 = u_0\},$$

is trivial, then both phase conditions (2.26) and (2.27) satisfy Hypothesis 2.8.

Proof. It is well known (see [6, Thm. 4.3.4]) that

$$\dim(\mathcal{O}(u_0)) = \dim(T_{u_0}\mathcal{O}(u_0)) = \dim(G) - \dim(\mathrm{Stab}(u_0)).$$

Let d be the dimension of G. Then we can choose elements $g_i \in \mathcal{A}$, $i = 1, \dots, d$, such that $S_i = [a_\gamma(\mathbb{1})u_0]g_i$ form a basis of $T_{u_0}\mathcal{O}(u_0)$. By a direct calculation, we have for $\lambda, \mu \in \mathcal{A}$

$$-[\psi_{1,v}(u_0, \mathbb{1})a_\gamma(u_0)\lambda]\mu = \langle a_\gamma(\mathbb{1})u_0\mu, a_\gamma(\mathbb{1})u_0\lambda \rangle. \tag{2.28}$$

Therefore, $\psi_{1,v}(u_0, \mathbb{1})a_\gamma(u_0)\lambda : \mathcal{A} \mapsto \mathcal{A}^*$ is nonsingular if and only if the $(d \times d)$-matrix with entries $\langle S_i, S_j \rangle$ is nonsingular. Clearly this holds if and only if the S_i are linearly independent. For the second phase condition, (2.27), note that we obtain the same expression, (2.28). ■

Remark 2.11. If the inner product is G-invariant, then $e_1(\gamma) = e_2(\gamma^{-1})$ and the two minimization problems are equivalent. Moreover, the two necessary conditions, (2.26) and (2.27),

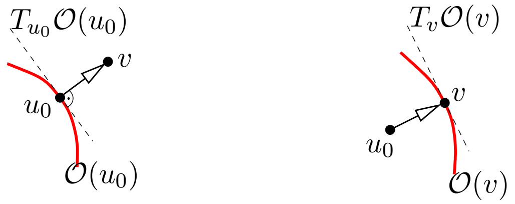


Figure 2.2. *Minimizing the distance of v to $\mathcal{O}(u_0)$ (left) or of u_0 to $\mathcal{O}(v)$ (right).*

are identical since we have $\langle a_\gamma(\mathbb{1})v, v\rangle = 0$ for $v \in Y$. The last equation follows by differentiating $\langle a(\gamma)v, a(\gamma)v\rangle = \langle v, v\rangle$ at $\gamma = \mathbb{1}$.

In equivariant bifurcation theory, interesting phenomena arise when the isotropy group of some relative equilibrium is nontrivial; see [10], [6]. For an initial value problem with some "generic" u_0, however, it seems reasonable to assume a trivial isotropy subgroup.

The remark applies to the parabolic system (2.23)–(2.25) with spaces (2.3). In this case the phase conditions (2.26) and (2.27) yield the integral constraint

$$0 = \psi(v) = \int_{-\infty}^{\infty} u_{0,x}^T(u_0 - v)\, dx = -\int_{-\infty}^{\infty} u_{0,x}^T v\, dx. \tag{2.29}$$

Hypothesis 2.8 requires the map $\lambda \to \lambda \int_{-\infty}^{\infty} u_{0,x}^T u_{0,x}\, dx$ to be nonsingular, which is satisfied if u_0 is nonconstant.

The phase conditions developed so far depend on the initial value u_0 and seem to be useful only for short times; see Theorem 2.9. During numerical computations one could update the phase condition by using $v(t_1), v(t_2), \dots$ at later times instead of u_0.

A condition that is applicable in a more global sense is minimizing the temporal change of v, i.e.,

$$|v_t|^2 = |F(v) - a(\gamma^{-1})a_\gamma(\gamma)v\lambda|^2. \tag{2.30}$$

This is a d-dimensional least squares problem in $\lambda \in \mathcal{A}_\gamma$. We introduce the operators

$$S(v,\gamma) = a(\gamma^{-1})a_\gamma(\gamma)v : \mathcal{A}_\gamma \mapsto X \tag{2.31}$$

and $S^*(v,\gamma) : X \mapsto \mathcal{A}_\gamma^*$ by

$$[S^*(v,\gamma)u]\lambda = \langle S(v,\gamma)\lambda, u\rangle \quad \text{for } \lambda \in \mathcal{A}_\gamma. \tag{2.32}$$

If the stabilizer of $a(\gamma)v$ is trivial, then $S(v,\gamma)$ is one to one and $S^*S(v,\gamma) \in L(\mathcal{A}_\gamma, \mathcal{A}_\gamma^*)$ is nonsingular. Therefore, (2.30) has a unique minimizer given by the solution of the linear $(d \times d)$-system

$$\psi_{\min}(v,\gamma,\lambda) := (S^*S)(v,\gamma)\lambda - S^*(v,\gamma)F(v) = 0. \tag{2.33}$$

Note that in contrast to (2.25) this phase condition depends also on the derivative $\gamma_t = \lambda$ so that Theorem 2.9 does not apply. Nevertheless, a given solution $u(t)$, $t \in [0,T)$, of (1.1) may be written as $u(t) = a(\gamma(t))v(t)$ with v satisfying (2.20) and (γ,λ) satisfying (2.33) if we determine $\gamma(t)$ from the following initial value problem on G:

$$\gamma_t(t) = [(S^*S)^{-1}S^*](a(\gamma^{-1})u(t),\gamma)\ a(\gamma^{-1})F(u(t)), \quad \gamma(0) = \mathbb{1}. \tag{2.34}$$

In order to ensure a unique solution of this problem we need more regularity for $F(u(t))$, $t \in [0,T)$, than in Theorem 2.9, such that the right-hand side of (2.34) is continuous in (γ,t) and locally Lipschitz in γ.

More details on the implementation of the phase condition (2.33) will be given in the following sections. The condition turns out to be particularly useful near relative equilibria of (1.1), where we expect v_t to tend to zero.

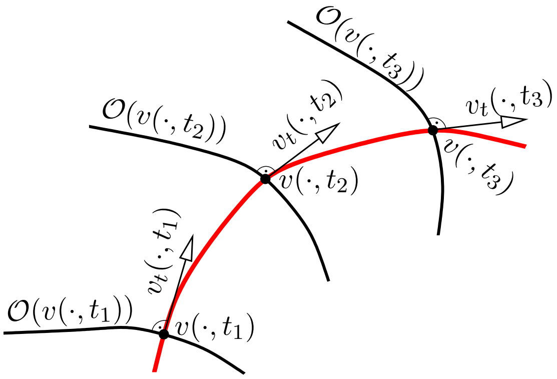


Figure 2.3. *Orthogonality of time and group orbit at successive times.*

A final alternative is to require that, at any time instance, v_t is orthogonal to the group orbit $\mathcal{O}(u) = \mathcal{O}(a(\gamma)v)$ at v (see Figure 2.3).

This leads to the condition

$$0 = \langle S(v, \mathbb{1})\mu, v_t\rangle \quad \forall \mu \in \mathcal{A}_\gamma.$$

Using the differential equation (2.20) we rewrite the phase condition as

$$\psi_{\text{orth}}(v, \gamma, \lambda) = S^*(v, \mathbb{1})S(v, \gamma)\lambda - S^*(v, \mathbb{1})F(v) = 0. \tag{2.35}$$

Note that this condition is identical with (2.33) if $a(\gamma^{-1})a_\gamma v$ does not depend on γ.

For example, it is true for the parabolic system (2.23)–(2.25) (and for most of the examples in sections 3 and 4). Conditions (2.33) and (2.35) both lead to the explicit formula

$$\langle v_x, v_x\rangle_{L^2}\lambda = \langle v_x, Av_{xx} + f(v, v_x)\rangle_{L^2} = \langle v_x, f(v, v_x)\rangle_{L^2}, \tag{2.36}$$

which works whenever the function v is nonconstant. Note that the last equality follows from $v \in H^2(\mathbb{R}, \mathbb{R}^m)$ if $A = A^T$.

2.4. Construction of spaces. In [29, Thm. 4.5] the authors set up a general principle for constructing spaces that satisfy the differentiability condition in Hypothesis 2.4. In the following proposition we slightly extend their result by constructing a sequence of nested spaces on which the group acts with increasing smoothness. We use the exponential map $\exp : \mathcal{A} \mapsto G$, cf. [6, Ch. 4.2].

Proposition 2.12. *Let $(X_0, ||\cdot||_0)$ be a Banach space and let $a : G \mapsto GL(X_0)$ be a homomorphism. Then*

$$X_1 = \left\{ u \in X_0 : ||u||_1 := \sup_{\gamma\in G} ||a(\gamma)u||_0 < \infty \right\}$$

is a Banach space with respect to the norm $||\cdot||_1$, and the operators $a(\gamma)_{|X_1}$ are isometries in $GL(X_1)$. Further, the space

$$X_2 = \{u \in X_1 : \gamma \mapsto a(\gamma)u \text{ is continuous in } G\}$$

is a closed subspace of $(X_1, ||\cdot||_1)$ *such that* $a(\gamma)_{|X_2} \in GL(X_2)$ *acts strongly continuously. Finally,*

$$X_3 = \{u \in X_2 : \gamma \mapsto a(\gamma)u \textit{ is continuously differentiable in } G\}$$

is a dense subspace of X_2 *and can be written as*

$$X_3 = \bigcap_{\lambda \in \mathcal{A}} D(\lambda), \tag{2.37}$$

where $D(\lambda)$ *is the domain of the infinitesimal generator of the* C^0*-semigroup* $a(\exp(\lambda t))$*,* $t \geq 0$*.*

Proof. If $u_n \in X_1$ is a Cauchy sequence with respect to $||\cdot||_1$, then $a(\gamma)u_n$ is a Cauchy sequence in X_0 for each $\gamma \in G$ and hence converges to some $v(\gamma) \in X_0$. By continuity of $a(\gamma)$ we have $v(\gamma) = a(\gamma)v(\mathbb{1})$, and using the Cauchy property again we obtain $||u_n - u||_1 \to 0$ for $u = v(\mathbb{1})$ as well as $u \in X_1$. The isometric property of $a(\gamma)_{|X_1}$ is obvious and the closedness of X_2 with respect to $||\cdot||_1$ is an easy exercise. The main result in [29, Thm. 4.5] states that $\bigcap_{\lambda \in \mathcal{A}} D(\lambda)$ is contained in X_3 and is a dense subspace of X_2. But the opposite inclusion $X_3 \subset \bigcap_{\lambda \in \mathcal{A}} D(\lambda)$ follows from the chain rule applied to $a(\exp(\lambda t))u$, $u \in X_3$, and this finishes the proof. ■

Remark 2.13. Under the assumptions of the proposition we can satisfy Hypothesis 2.4 by taking $X = X_2$ and $Y = X_3$. However, in the applications the right-hand side of (1.1) may contain differential operators that require an even smaller (but still dense) domain Y.

Example 2.14. For $\gamma \in G = \mathbb{R}^N$ consider the shift (see (2.9))

$$[a(\gamma)u](x) = u(x - \gamma), \quad x \in \mathbb{R}^N.$$

If we take $X_0 = C_b^0(\mathbb{R}^N, \mathbb{R}^m)$ (continuous bounded functions) with $||\cdot||_0$ as the sup-norm in Proposition 2.12, then we obtain $X_1 = X_0$, $||\cdot||_1 = ||\cdot||_0$, and the spaces of uniformly continuous functions $X_2 = C^0_{\mathrm{unif}}(\mathbb{R}^N, \mathbb{R}^m)$, $X_3 = C^1_{\mathrm{unif}}(\mathbb{R}^N, \mathbb{R}^m)$.

Another choice is that of locally uniform spaces as proposed in [20], [19]. Take a positive and integrable weight function $\eta \in C^1(\mathbb{R}^N, (0,\infty)) \cap L^1(\mathbb{R}^N)$ that satisfies $|\nabla \eta(x)| \leq C\eta(x)$ for all $x \in \mathbb{R}^N$. For $p \geq 1$ consider the weighted L^p space

$$X_0 = L^p_\eta(\mathbb{R}^N) = \{u \in L_{\mathrm{loc}}(\mathbb{R}^N, \mathbb{R}^m) : ||u||_{L^p_\eta} < \infty\}, \tag{2.38}$$

$$||u||_{L^p_\eta} = \left(\int_{\mathbb{R}^N} \eta(x)|u(x)|^p dx \right)^{1/p}. \tag{2.39}$$

From the estimate $\eta(x+\gamma) \leq e^{C|\gamma|}\eta(x)$ one finds that $a(\gamma) : X_0 \mapsto X_0$ is a bounded operator with bound $e^{\frac{C|\gamma|}{p}}$. The construction in Proposition 2.12 then yields the locally uniform spaces (using the notation from [20], [19])

$$X_0 \supset X_1 = \tilde{L}^p_{\mathrm{ul}}(\mathbb{R}^N) \supset X_2 = L^p_{\mathrm{ul}}(\mathbb{R}^N). \tag{2.40}$$

Here the norm $||u||_{L^p_{\mathrm{ul}}} = \sup_{\gamma \in \mathbb{R}^N} ||u(\cdot - \gamma)||_{L^p_\eta}$ in X_1 is stronger than $||\cdot||_0$ and all inclusions are strict. Finally, the intersection of the domains of the infinitesimal generators $\frac{\partial}{\partial x_j}$, $j = 1, \ldots, N$, leads to the weighted Sobolev space

$$X_3 = W^{1,p}_{\mathrm{ul}}(\mathbb{R}^N) = \left\{ u \in L^p_{\mathrm{ul}}(\mathbb{R}^N) : \frac{\partial u}{\partial x_j} \in L^p_{\mathrm{ul}}(\mathbb{R}^N), j = 1, \ldots, N \right\}. \tag{2.41}$$

3. Waves in one space dimension.

3.1. Relative equilibria. Following [10], [6] we define relative equilibria as solutions that stay in the group orbit of the initial value (see also [10] and [28] for further notions of a relative periodic orbit and meandering solutions).

Definition 3.1. *A solution $u(t)$, $t \in [0,T)$, of* (1.1) *is called a* relative equilibrium *if there exist $v \in Y$, $\gamma \in C^1([0,T),G)$ such that $\gamma(0) = \mathbb{1}$ and*

$$u(t) = a(\gamma(t))v, \quad 0 \le t < T. \tag{3.1}$$

In view of (2.20), (2.21) this implies the following equations:

$$0 = F(v) - S(v,\gamma(t))\lambda(t), \text{ where } \quad S(v,\gamma)\lambda = a(\gamma^{-1})a_\gamma(\gamma)v\lambda, \tag{3.2}$$

$$\gamma_t(t) = \lambda(t), \quad \gamma(0) = \mathbb{1}. \tag{3.3}$$

In the applications, we will frequently have relative equilibria for which the operator $S(\cdot,\gamma(t))\lambda(t) : Y \mapsto X$ is independent of t.

For example, a traveling wave

$$[u(t)](x) = v(x - \lambda t), \quad \gamma(t) = \lambda t, \tag{3.4}$$

is an equilibrium of system (2.23) with constant λ and a relative equilibrium of (2.2) (take any of the spaces from Example 2.14). Conversely, if $u(t) = a(\gamma(t))v$ is a relative equilibrium of (2.2), then with $\lambda = \gamma_t$ we have

$$0 = Av_{xx} + f(v, v_x) + v_x\lambda(t).$$

Taking the inner product with v_x and assuming that v is nonconstant, we conclude that $\lambda(t)$ is in fact time independent. Hence, traveling waves are the only nontrivial relative equilibria of (2.2).

As a second example consider the complex valued system

$$u_t = Au_{xx} + f(u, u_x), \quad x \in \mathbb{R}, \quad u(x,t) \in \mathbb{C}^m, \tag{3.5}$$

where $f : \mathbb{C}^{2m} \mapsto \mathbb{C}^m$ is assumed to be equivariant with respect to phase factors

$$f(e^{i\theta}u, e^{i\theta}v) = e^{i\theta}f(u,v), \ \theta \in S^1 = \mathbb{R}/2\pi\mathbb{Z}. \tag{3.6}$$

Well-known special cases are equations of *Ginzburg–Landau* type:

$$f(u,v) = au + bu|u|^2 + cu|u|^4, \ u \in \mathbb{C}, \quad a,b,c \in \mathbb{C}. \tag{3.7}$$

In this case the Lie group is $G = S^1 \times \mathbb{R}$ with

$$(\theta_1, \tau_1) \circ (\theta_2, \tau_2) = (\theta_1 + \theta_2, \tau_1 + \tau_2) \tag{3.8}$$

and the action is given by

$$[a(\theta,\tau)v](x) = e^{-i\theta}v(x - \tau). \tag{3.9}$$

System (2.20), (2.21) now reads

$$v_t = Av_{xx} + f(v, v_x) + iv\lambda_1 + v_x\lambda_2, \qquad v(0) = u_0, \tag{3.10}$$

$$\theta_t = \lambda_1,\ \tau_t = \lambda_2, \qquad \theta(0) = 0, \tau(0) = 0. \tag{3.11}$$

The phase condition (2.26) has the form

$$\psi_1(v, \theta, \tau) = (\langle iu_0, v - u_0\rangle, \langle u_{0,x}, v - u_0\rangle) = (0, 0), \tag{3.12}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in the real system of doubled dimension, i.e.,

$$\langle u + iv, w + iz\rangle_{L_2} = \int_{\mathbb{R}} u^T w + v^T z\ dx. \tag{3.13}$$

Hypothesis 2.8 is satisfied if the functions iu_0 and u_{0x} are linearly independent over $\mathbb{R}$. Relative equilibria of (3.5) are rotating waves

$$u(x, t) = e^{-i\lambda_1 t} v(x - \lambda_2 t), \quad x \in \mathbb{R},\ t \in \mathbb{R}, \tag{3.14}$$

where v is in one of the spaces $H^2(\mathbb{R}, \mathbb{C}^m)$, $C^2_{\text{unif}}(\mathbb{R}, \mathbb{C}^m)$, or $H^2_{\text{ul}}(\mathbb{R}, \mathbb{C}^m) = W^{2,2}_{\text{ul}}(\mathbb{R}, \mathbb{C}^m)$; cf. Example 2.14. Similarly to the previous example, we obtain that these are the only relative equilibria for which iv and v_x are linearly independent.

3.2. Numerical computations. For the discretization in time and space we consider a system of the form (2.2) that is equivariant under a Lie group of dimension d and that—with a proper choice of coordinates in G and TG—leads to a PDAE (2.20), (2.21) of the form

$$v_t = F(v) - S(v)\lambda, \quad S(v)\lambda = \sum_{j=1}^{d} S_j(v)\lambda_j, \qquad v(0) = u_0, \tag{3.15}$$

$$\gamma_t = \lambda, \qquad \gamma(0) = 0. \tag{3.16}$$

Here F is given by (2.2) and the S_j are linear differential operators of order ≤ 1 with bounded continuous coefficients

$$S_j(v) = B_j(x)v_x + C_j(x)v, \quad B_j, C_j \in C^0_b(\mathbb{R}, \mathbb{R}^{m,m}).$$

For the phase condition we use the orthogonality constraint (2.35), which in this case is the same as (2.33):

$$\psi_{\text{orth}}(v, \lambda) = \left(\left\langle S_\nu(v), \sum_{j=1}^{d} S_j(v)\lambda_j - F(v)\right\rangle\right)_{1\leq\nu\leq d} = 0, \tag{3.17}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in either L^2 or L^2_η.

We choose a step size Δt in time and an equidistant spatial grid

$$J = \{x_j = j\Delta x : 0 \leq j \leq M\}, \quad \Delta x = \frac{x_+ - x_-}{M},$$

where $[x_-,x_+]$ is some large interval. At time $t_n = n\Delta t$ we compute the approximations $\gamma^n,\lambda^n \in \mathbb{R}^d$ and $v^n : J \mapsto \mathbb{R}^m$ by a half-explicit Euler method, i.e., a method that is explicit in the state variable v^n but implicit in the algebraic variable λ^n (see [15], [16] for such methods):

$$\frac{1}{\Delta t}(v^{n+1}-v^n) = F_{\Delta x}(v^n) - S_{\Delta x}(v^n)\lambda^{n+1}, \tag{3.18}$$

$$\frac{1}{\Delta t}(\gamma^{n+1}-\gamma^n) = \frac{1}{2}(\lambda^{n+1}+\lambda^n), \tag{3.19}$$

$$0 = \left(\left\langle S_{\nu,\Delta x}(v^n), \sum_{j=1}^{d} S_{j,\Delta x}(v^n)\lambda_j^{n+1} - F_{\Delta x}(v^n)\right\rangle\right)_{1\le\nu\le d}. \tag{3.20}$$

Here $F_{\Delta x}$ and $S_{\Delta x}$ are standard finite difference approximations

$$F_{\Delta x}(v) = D_+D_-v + f(v, D_0 v), \quad S_{j,\Delta x}(v) = B_j D_0 v + C_j v,$$

where $D_\pm$ denote forward/backward and $D_0 = \frac{1}{2}(D_+ + D_-)$ centered difference quotients. In any time step one first solves the linear $(d\times d)$-system (3.20) for λ^{n+1} and then determines v^{n+1},γ^{n+1} from (3.18), (3.19). Standard stability restrictions such as $\Delta t \le \frac{1}{2}(\Delta x)^2$ are taken into account.

Equation (3.18) has to be completed by boundary conditions. We choose Neumann and projection boundary conditions.

For traveling waves (more generally, relative equilibria), projection boundary conditions are commonly used as asymptotic boundary conditions at $x_\pm$ in order to have higher order approximations of wave form and speed (see [4], [26]). We adapt them to the time-dependent case as follows. Assume that the limits $\lim_{x\to\pm\infty} C_j(x) = C_{j,\pm}$ exist and the solution satisfies

$$\lim_{x\to\pm\infty} v(x,t) = w_\pm, \ \lim_{x\to\pm\infty} v_x(x,t) = 0, \ f(w_\pm,0) - \sum_{j=1}^{d} C_{j,\pm} w_\pm = 0. \tag{3.21}$$

The idea behind projection boundary conditions is to control the growing, resp., decaying, spatial modes obtained by linearizing at $\pm\infty$. We write the conditions in the form

$$V_\pm(v(x_\pm) - w_\pm) + W_\pm v_x(x_\pm) = 0, \tag{3.22}$$

where $V_\pm, W_\pm \in \mathbb{R}^{m,m}$. These matrices can be obtained by setting $W_\pm = Z_\pm A$, $V_\pm = -\Lambda_\pm^{-1} Z_\pm C_\pm$, where $\Lambda_+,\Lambda_- \in \mathbb{R}^{m,m}$ have only eigenvalues with real part positive, resp., negative, and where $Z_\pm \in \mathbb{R}^{m,m}$ form a corresponding invariant subspace of the "left quadratic eigenvalue problem":

$$\Lambda_\pm^2 Z_\pm A + \Lambda_\pm Z_\pm B_\pm + Z_\pm C_\pm = 0, \tag{3.23}$$

$$B_\pm = D_2 f(w_\pm,0) + \sum_{j=1}^{d} B_{j,\pm}\lambda_j, \quad C_\pm = D_1 f(w_\pm,0) + \sum_{j=1}^{d} C_{j,\pm}\lambda_j. \tag{3.24}$$

In the nth time step one has to set $\lambda = \lambda^n$ in (3.24), and so the projection boundary conditions (3.22) depend on time.

3.3. Numerical examples. In the following we test our method on several well-known examples of parabolic systems that show traveling or rotating waves.

3.3.1. Nagumo wave. The Nagumo equation [21], [22]

$$u_t = u_{xx} + u(1-u)(u-\alpha),\ u(x,0) = u_0(x), \quad \alpha \in \left(0, \frac{1}{2}\right) \tag{3.25}$$

has an explicit traveling wave solution $u(x,t) = \bar{v}(x-ct)$ given by

$$\bar{v}(x) = \frac{1}{1+\exp(-\frac{x}{\sqrt{2}})}, \quad c = -\sqrt{2}\,\left(\frac{1}{2}-\alpha\right).$$

For the numerical computations, we use parameters $\alpha = \frac{1}{4}$, $J = [-30, 30]$, $\Delta x = 0.1$, $\Delta t < \frac{1}{2}\Delta x^2$, and Neumann boundary conditions.

In In Figure 3.1 we compare the time evolution of a piecewise linear initial profile for the unmodified equation with its frozen counterpart computed from (3.18)–(3.20). The frozen profile stabilizes after a short time, and the parameter $\lambda(t)$ converges to a fixed value $\lambda_\infty = -0.353555$, which is in good agreement with the velocity of the exact solution on $\mathbb{R}$. In contrast to this, the solution of the original equation becomes constant when the wave reaches the left

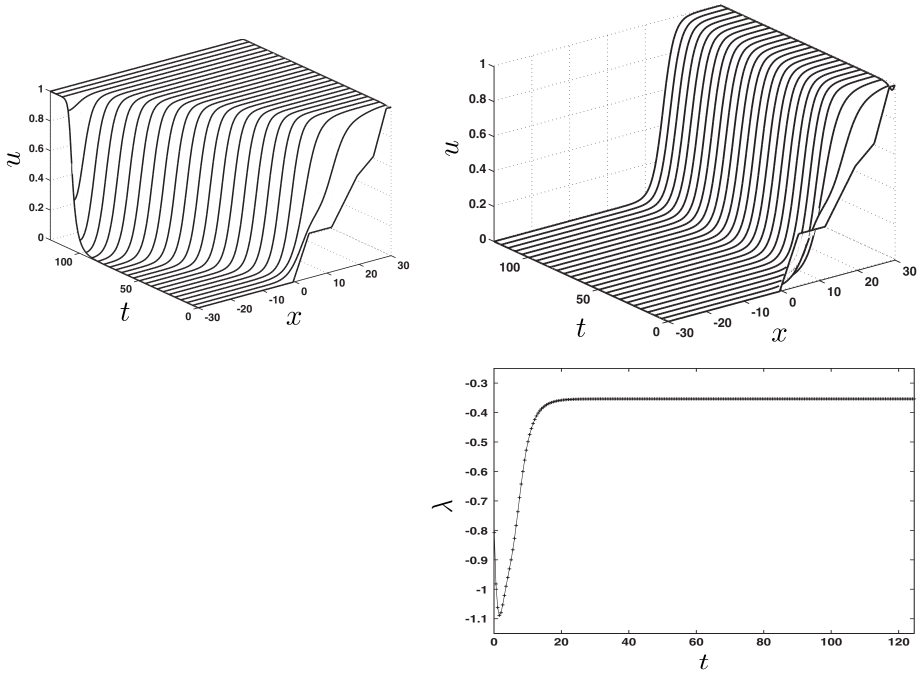
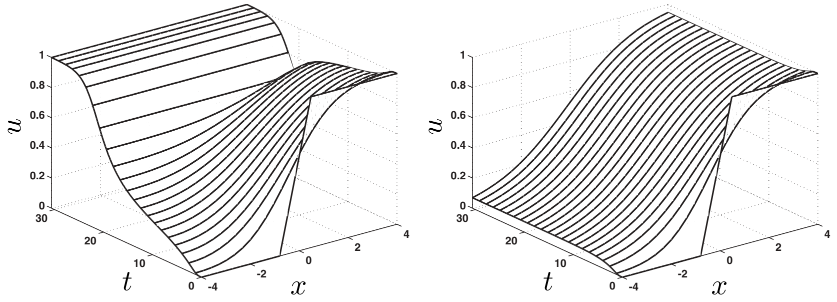


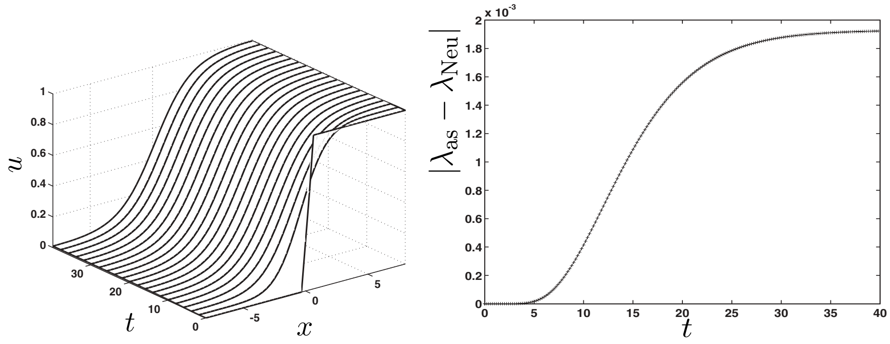
Figure 3.1. *Traveling versus frozen Nagumo wave and evolution of velocity $\lambda(t)$.*



(a) Neumann boundary conditions

(b) Asymptotic boundary conditions

Figure 3.2. *Frozen Nagumo wave on* $J = [-4, 4]$.



(a) Wave profile

(b) Difference to asymptotic boundary conditions in λ

Figure 3.3. *Frozen Nagumo wave on* $J = [-8, 8]$*, Neumann boundary conditions.*

boundary. The reason is, of course, that the constants 0, 1, and α are the only solutions of the stationary boundary value problem on the finite interval. While the evolution problems (3.25) and (3.15), (3.16) are equivalent on the whole real line (cf. Theorem 2.6) they become different when truncated to a finite interval. Figure 3.2 we compare two frozen equations with different boundary conditions (Neumann and asymptotic) on a rather short interval of $J = [-4, 4]$. While asymptotic boundary conditions still admit a stationary profile close to the original wave, Neumann boundary conditions allow only constant stationary profiles as solutions of the corresponding frozen equation in the limit $t \to \infty$.

On the larger interval $J = [-8, 8]$, Neumann boundary conditions are acceptable again; see Figure 3.3. The solutions for both boundary conditions differ by an amount of the order 10^{-3}, which cannot be seen in the scale of Figure 3.3(a). Therefore, we show in Figure 3.3(b) the difference in λ as a function of time. Summarizing, the pictures demonstrate that the advantages of projection boundary conditions for the computation of stationary profiles carry over to the time-dependent case.

3.3.2. FitzHugh–Nagumo wave. A well-known two-component system with traveling wave solutions is given by the FitzHugh–Nagumo equations [21], [22]

$$\begin{aligned} V_t &= \Delta V + V - \frac{1}{3}V^3 - R, \\ R_t &= \phi(V + a - bR) \end{aligned}$$

with parameters $a = 0.7$, $b = 0.8$, $\phi = 0.08$. As in the Nagumo case, the equation is equivariant with respect to translation. The numerical parameters are $J = [0, 130]$, $\Delta x = 0.5$, and $\Delta t = 0.01$ with Neumann boundary conditions.

In Figure 3.4 the time evolution of the V-component of a given initial profile (R has been set initially to the stationary value $\bar{R} = -0.62426$) is shown. The initial hump splits into two traveling components, and after some time only the left moving pulse exists and leaves the computational window. Figure 3.5 shows the solution of the frozen equation starting from the same profile. As before, the initial profile splits into a left and a right traveling pulse. When the right moving solution has left J, the remaining pulse stabilizes and takes the shape of the well-known stable pulse (see [21]). The parameter $\lambda(t)$ converges after a transition phase to $\lambda_\infty = -0.816848$. As we see in this example, our method can only freeze one wave at a time. Which one is selected depends on the type of phase condition used.

In Figure 3.5(a) the adaptive phase condition (3.17) is used and the left moving pulse is frozen. If we use the fixed phase condition (2.26), as shown in Figure 3.5(b), the right moving pulse stabilizes at the position of the initial hump.

3.3.3. The complex Ginzburg–Landau equation. We consider a special normalization of the complex Ginzburg–Landau equation discussed in [19]:

$$u_t = (1 + i\alpha)(u_{xx} - (1 + i\omega)^2 u + (1 + i\omega)(2 + i\omega)|u|^2 u), \quad u = u_1 + iu_2. \tag{3.26}$$

Here α and ω are real parameters. As described in section 3.1, this equation is equivariant under the action of the symmetry group $G = S^1 \times \mathbb{R}$.

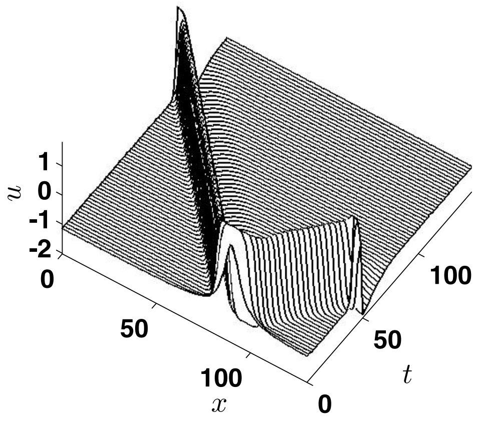
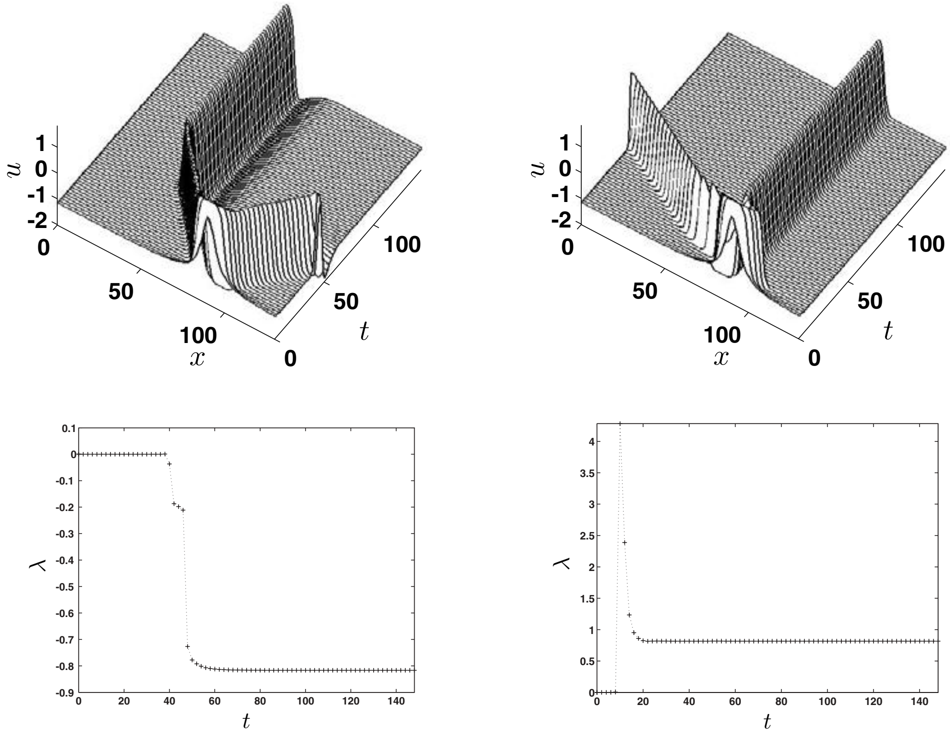


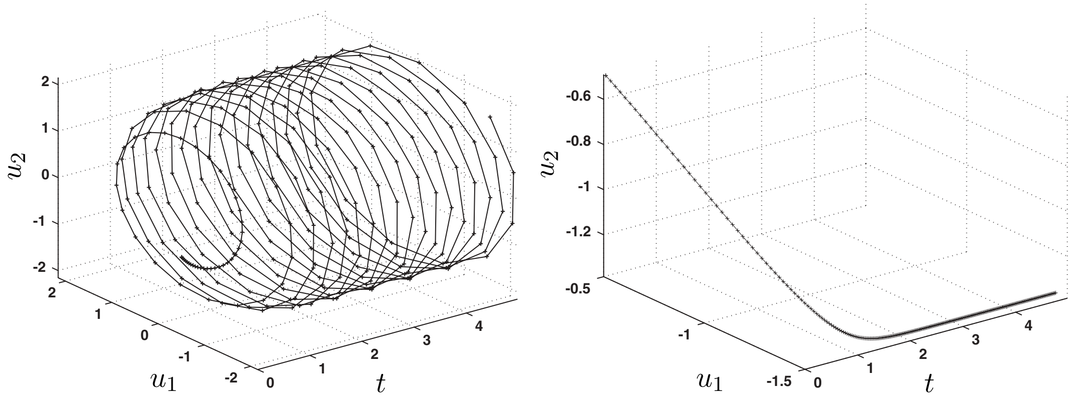
Figure 3.4. *Fitzhugh–Nagumo traveling wave.*



(a) Phase condition (3.17)

(b) Phase condition (3.12)

Figure 3.5. *FHN, frozen wave.*



(a) Rotating solution

(b) Frozen solution

Figure 3.6. *Complex Ginzburg–Landau system, rotating vs. frozen solution.*

One finds rotating wave solutions of the form $u(x,t) = e^{i\phi t}u_0$ with a profile u_0, which is constant in x. Inserting this ansatz into (3.26), one obtains, for the absolute value of the solution u_0 and for the angular velocity ϕ, the formulas

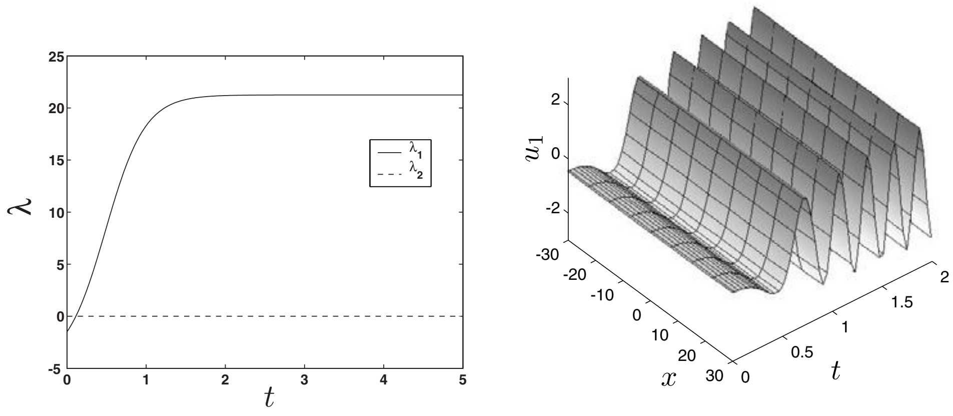
$$|u_0|^2 = \frac{\omega^2 + 2\alpha\omega - 1}{\omega^2 + 3\alpha\omega - 2}, \quad \phi = (\alpha(2 - \omega^2) + 3\omega)|u_0|^2 + \alpha(\omega^2 - 1) - 2\omega.$$

For the numerical computations we choose parameters $\omega = -2$, $\alpha = \frac{1}{4}$, $J = [-30, 30]$, $\Delta x = 0.5$, $\Delta t = 0.001$, and Neumann boundary conditions. We start from an initial profile, which consists of small Gaussian pulses for the components u_1 and u_2.

In Figure 3.6 the time evolution of the point $u(0,t)$ of the solution for the rotating and the frozen system are compared. The frozen solution stabilizes after some time at a fixed value whereas the solution of the original system continues rotating.

As shown in Figure 3.7(a) the parameter $\lambda_1(t) = \dot{\theta}(t)$ converges to the exact angular velocity $\phi = 21.25$, and the translational speed $\lambda_2(t) = \dot{\tau}(t)$ stays at zero, as expected. The evolution of the whole profile of the u_1 component of the rotating solution is depicted in Figure 3.7(b).

Figure 3.8 demonstrates the advantage of the half-explicit scheme for the frozen system over the explicit scheme for the original system. We compare the discretized analogue of the normalized L^2 norm of the solution $\|u(\cdot,t)\|_n = \sqrt{\frac{1}{|J|}\int_J \|u(x,t)\|^2 \, dx}$ for the frozen and the rotating system. Since the parameter λ is computed implicitly from the phase condition (3.20), the norm $\|u(\cdot,t)\|_n$ converges for the frozen system to the exact value of 2, whereas in the rotating system the norm is overestimated due to the use of the explicit Euler method.



(a) Evolution of $\lambda_1 = \dot{\theta}$, $\lambda_2 = \dot{\tau}$

(b) Evolution of the initial profile

Figure 3.7. *Complex Ginzburg–Landau equation.*

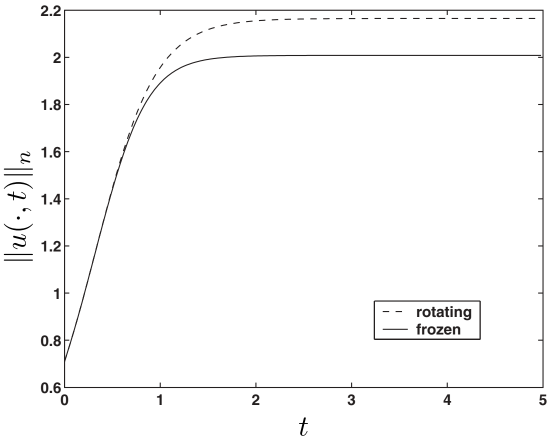


Figure 3.8. $\|u(\cdot,t)\|_n$, *rotating versus frozen system.*

4. Spiral waves in two dimensions.

4.1. The PDAE for Euclidean symmetry. Consider the semilinear parabolic system (2.4), i.e.,

$$u_t = A\Delta u + f(u), \quad x \in \mathbb{R}^2, t \geq 0, \quad u(\cdot,0) = u_0, \tag{4.1}$$

where $A \in \mathbb{R}^{m,m}$ is positive definite and $f \in C^\infty(\mathbb{R}^m, \mathbb{R}^m)$. The system is equivariant with respect to action (2.12) of the Euclidean group $SE(2)$ and satisfies (2.1) with the spaces (see [28], [29])

$$Y = C^2_{\text{unif}}(\mathbb{R}^2, \mathbb{R}^m), \quad X = C^0_{\text{unif}}(\mathbb{R}^2, \mathbb{R}^m). \tag{4.2}$$

It is proved in [20] that this is also true for certain cubic nonlinearities for the uniformly local spaces (cf. Example 2.14)

$$Y = H^2_{\mathrm{ul}}(\mathbb{R}^2, \mathbb{R}^m), \quad X = L^2_{\mathrm{ul}}(\mathbb{R}^2, \mathbb{R}^m). \tag{4.3}$$

Formally differentiating action (2.12) with respect to $\gamma = (\theta, b) \in S^1 \ltimes \mathbb{R}^2$ yields the expression

$$S(v,\theta)\lambda := a(\gamma^{-1})\left[a_\gamma(\gamma) v\lambda\right] = -v_x\left[\varrho_{\frac{\pi}{2}} x\lambda_1 + \varrho_{-\theta}\begin{pmatrix}\lambda_2\\ \lambda_3\end{pmatrix}\right] \tag{4.4}$$

for $\lambda_1 \in S^1$, $\lambda_2, \lambda_3 \in \mathbb{R}$.

This formula can be shown rigorously, and Hypothesis 2.4 is satisfied if the function v lies in the space

$$\tilde{Y} = \{v \in Y : Pv \in X\},$$

where

$$(Pv)(x) = v_x(x)\varrho_{\frac{\pi}{2}} x = -x_2 v_{x_1} + x_1 v_{x_2}. \tag{4.5}$$

Note that this follows from Proposition 2.12 since the domain of P corresponds to $D(\lambda_1)$ in (2.37) and since Y is contained in the domain of the other two infinitesimal generators $\frac{\partial}{\partial x_1}$ and $\frac{\partial}{\partial x_2}$.

System (2.20) is of the form

$$v_t = A\Delta v + f(v) - S(v,\theta)\lambda, \quad v(0) = u_0, \tag{4.6}$$

with $S(v,\theta)$ defined in (4.4), and (2.21) reads

$$\theta_t = \lambda_1, \quad b_t = \begin{pmatrix}\lambda_2\\ \lambda_3\end{pmatrix}, \quad \theta(0) = 0, \quad b(0) = 0. \tag{4.7}$$

In contrast to (3.15), the forcing term on the right-hand side of (4.6) depends on the group variable θ. We can eliminate this dependence by choosing new coordinates (θ, α) on G and μ on $\mathcal{A}$ as follows:

$$\alpha = \varrho_{-\theta} b, \quad \mu_1 = \lambda_1, \quad \begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix} = \varrho_{-\theta}\begin{pmatrix}\lambda_2\\ \lambda_3\end{pmatrix}.$$

This transforms (4.6), (4.7) into

$$v_t = A\Delta v + f(v) - S(v)\mu, \quad v(0) = u_0, \quad S(v)\mu = -v_x\left[\varrho_{\frac{\pi}{2}} x\mu_1 + \begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\right], \tag{4.8}$$

$$\theta_t = \mu_1, \quad \alpha_t = \mu_1 \varrho_{\frac{\pi}{2}}\alpha + \begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}, \quad \theta(0) = 0, \quad \alpha(0) = 0. \tag{4.9}$$

Note that the second equation in (4.9) is no longer trivial but describes, in the case of a constant μ, a rotation on a circle of radius $|\mu_1|$ about the center $\frac{1}{\mu_1}\left(\begin{smallmatrix}-\mu_3\\ \mu_2\end{smallmatrix}\right)$. In this version the two phase conditions (2.33) and (2.35) coincide.

Both systems (4.6) and (4.8) introduce a convection term Pv, which becomes large on large domains. The numerical discretization will take this into account; see section 4.2 below.

We write system (4.6) in polar coordinates, which are particularly well suited for spiral waves. With $w(r,\varphi) = v(r\cos\varphi, r\sin\varphi)$ we obtain

$$w_t = A\Delta_{r,\varphi} w + f(w) + \lambda_1 w_\varphi + \begin{pmatrix} w_r & \frac{1}{r} w_\varphi \end{pmatrix} \varrho_{-\varphi-\theta} \begin{pmatrix} \lambda_2 \\ \lambda_3 \end{pmatrix} \tag{4.10}$$

with the Laplacian $\Delta_{r,\varphi} w = w_{rr} + \frac{1}{r} w_r + \frac{1}{r^2} w_{\varphi\varphi}$. In the numerical experiments below we use a rectangular grid for (4.10), which corresponds to a polar grid for the original equation, (4.6). The numerical experiments show that the influence of the geometry of the domain is much stronger for the frozen system. Using a cartesian grid on a rectangular domain shows strong negative effects of the boundary. In particular, we were not able to freeze nonlocalized spirals in this situation.

An appropriate inner product that is continuous with respect to the topology of the space X in both cases (4.2), (4.3) (cf. section 2.3) is

$$\langle u, v\rangle_\eta = \int_{\mathbb{R}^2} u(x)^T v(x) \eta(x) dx, \tag{4.11}$$

where the weight function satisfies (see Example 2.14)

$$\eta \in C^1(\mathbb{R}^2, (0,\infty)) \cap L^1(\mathbb{R}^2), \quad |\nabla\eta(x)| \le C\eta(x), \quad x \in \mathbb{R}^2. \tag{4.12}$$

In some numerical examples below the choice of inner product actually makes a difference because it enters into the phase condition.

4.2. Numerical method. We consider the polar system (4.10) on a rectangle $[0,R]\times[0,2\pi)$ and use periodic boundary conditions in the φ-direction and Neumann boundary conditions $v_r = 0$ at $r = R$. So far we have not yet set up appropriate projection boundary conditions that generalize the one-dimensional case, (3.22). These will require us to solve a linearized exterior boundary value problem that we expect to be quite expensive. Note that in [9] a simple type of mixed boundary condition is proposed in order to create spiral solutions for a scalar equation.

For the discretization we choose a rectangular grid on $[0,R] \times [0,2\pi)$ with step sizes Δr and $\Delta\varphi$. Second order derivatives v_{rr} and $v_{\varphi\varphi}$ are replaced by centered difference quotients; at the origin we use the standard cartesian five-point formula. For the examples in sections 4.3.1 and 4.3.2 below we also use centered differences for the first derivatives. However, for the last example the artificial convection introduced in (4.10) dominates diffusion so that it was necessary to use an upwind-downwind scheme (see 4.3.3 for details) Time is again discretized by a half-explicit Euler method as in (3.18)–(3.20) with $S_{\Delta x}(v^n)\lambda^{n+1}$ replaced by the corresponding discretization of the forcing terms in (4.10). Stability restrictions caused by the use of the half-explicit method,

$$\Delta t \le C \min\{(\Delta r)^2, (\Delta\varphi)^2\},$$

are taken into account. Finally, integrals are replaced by trapezoidal sums.

4.3. Examples. The following computations are performed with a version of Barkley's code `ezspiral` [1] that has been adapted to the discretization described in the previous section. In all examples we used the phase condition ψ_{min} from (2.33), which minimizes the temporal change of v. In fact, the numerical values showed no substantial difference between ψ_{min} and the phase condition ψ_{orth} from (2.35), which guarantees orthogonality of time and group orbit.

As a general remark concerning all of our computations we note that the λ-values after one time step jump to a consistent value $\lambda = \lambda^1$ according to the algebraic condition (3.20). This is the normal behavior for DAEs, and our λ-plots start after this first step.

Movies are linked to some of the figures for better illustration and their color code is identical to the code used in the corresponding figures. The time development of the λ variables is shown in an extra diagram and the behavior of the group variables θ and b is indicated by a black bar and a white trace.

4.3.1. $(\lambda - \omega)$-system. Our first example is a $(\lambda - \omega)$-system [18] in the complex form

$$u_t = \Delta u + (\lambda(|u|) + i\omega(|u|))u, \quad u(x,t) \in \mathbb{C}, \tag{4.13}$$

where λ and ω are functions of $|u|$. We take $\lambda(|u|) = 1 - |u|^2$, $\omega(|u|) = -|u|^2$ for which rigidly rotating waves are known to exist [18]. As shown in section 4.1, (4.13) is equivariant with respect to the action of the group $SE(2)$ defined in (2.12). We solve the frozen system (4.10), (4.7) together with the phase condition defined by ψ_{min} using the L^2 inner product. The numerical parameters are $R = 50$, $\Delta r = 0.5$, $\Delta\varphi = \frac{\pi}{40}$, $\Delta t = 1.5421 \cdot 10^{-4}$. Starting from the initial function

$$u_0(r,\varphi) = \frac{r}{R}(\cos(\varphi) + i\sin(\varphi)), \tag{4.14}$$

a counterclockwise rotating spiral develops.

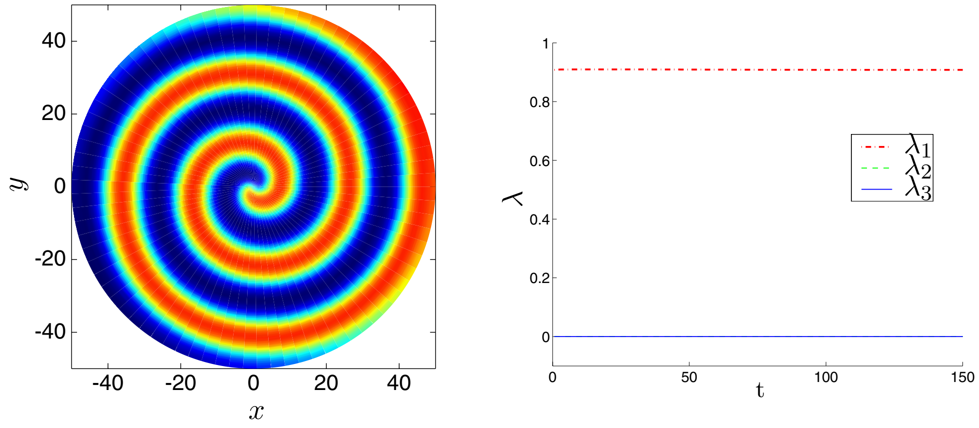
In Figure 4.1(a) a color scale plot of a snapshot of the real part of the spiral solution is shown at a fixed time instance.

In Figure 4.2 the time evolution of a slice along the x-axis of the spiral is compared for the rotating and the frozen system. As an initial condition we chose the centered spiral shown in Figure 4.1. Figure 4.1(b) shows the corresponding evolution of the λ parameters of the frozen system. The rotational velocity λ_1 instantly jumps to a fixed value of $\bar{\lambda} \approx 0.9$, while the translational speeds λ_2 and λ_3 remain at zero.

Figure 4.3 compares the time evolution for the frozen and the nonfrozen system starting at the initial value u_0 defined in (4.14) far away from relative equilibria. In both cases, eventually a spiral of the same shape develops (faster for the rotating wave than for the frozen wave).

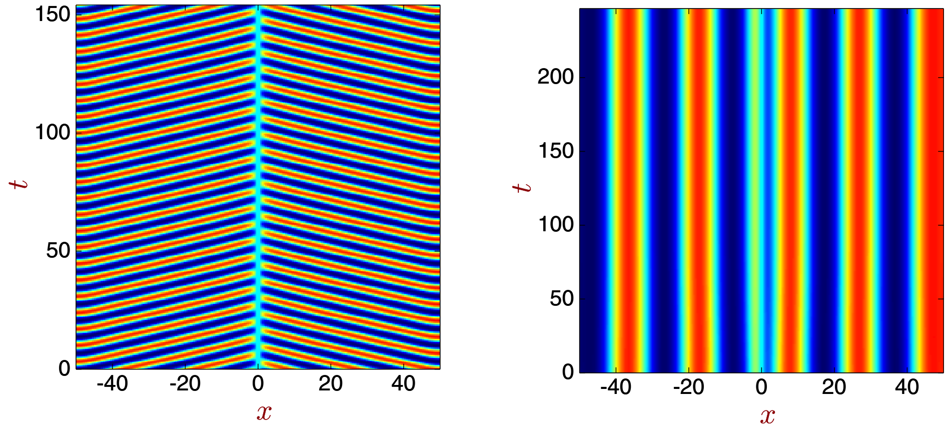
If we start with a slightly shifted spiral, we get the results shown in Figure 4.4. Near the boundary the spiral broadens slightly, an effect that decreases as the size of the domain increases. The (λ_2, λ_3)-variables in 4.4(b) show that the center of the spiral now rotates on a circle.

In fact, the system (4.13) shows equivariance with respect to the larger group $SE(2) \times S^1$. Denoting group elements by $\gamma = (\theta, b, \phi) \in (S^1 \ltimes \mathbb{R}^2) \times S^1$, we see that the action on real



(a) Spiral solution (b) Evolution of $\lambda(t)$

Figure 4.1. *$(\lambda-\omega)$-system.*

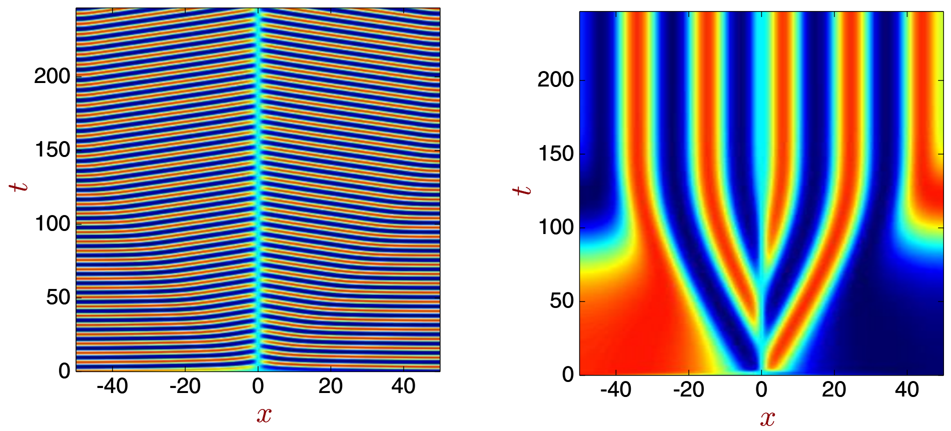


(a) Rotating spiral (b) Frozen spiral

Figure 4.2. *Time evolution of a slice for the rotating and the frozen spirals for the $(\lambda-\omega)$-system. Clicking on the above image shows the associated movie.*

valued functions and the forcing term in (2.31) are given by

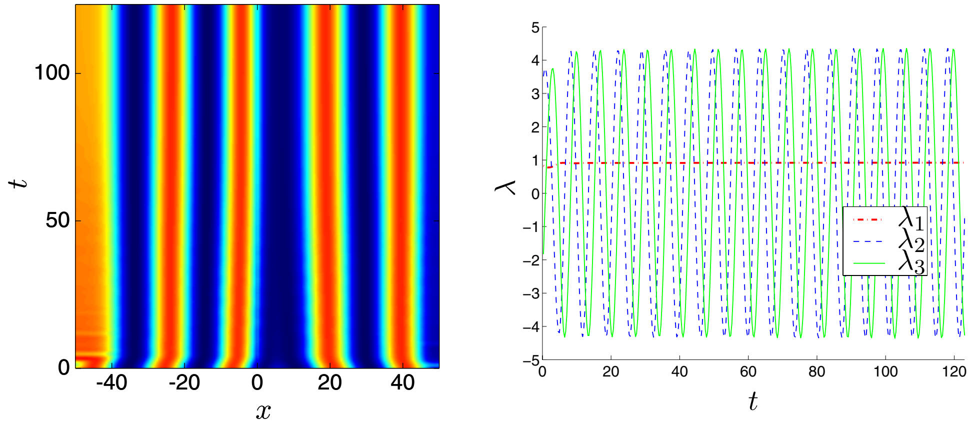
$$\begin{aligned}[a(\gamma)v](x) &= \varrho_{-\phi}v(\varrho_{-\theta}(x-b)),\\ S(v,\gamma)\lambda &= -v_x\left[\varrho_{\frac{\pi}{2}}x\lambda_1+\varrho_{-\theta}\begin{pmatrix}\lambda_2\\ \lambda_3\end{pmatrix}\right]-\varrho_{\frac{\pi}{2}}v\ \lambda_4.\end{aligned} \tag{4.15}$$



(a) Rotating spiral

(b) Frozen spiral

Figure 4.3. *Time evolution of a slice for the $(\lambda-\omega)$-system started far away from the spiral wave. Clicking on the above image shows the associated movie.*



(a) Frozen spiral

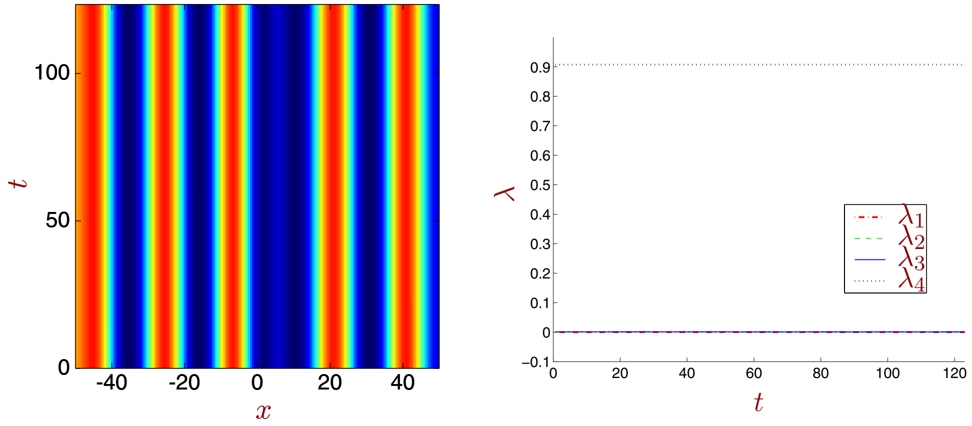
(b) Evolution of $\lambda(t)$

Figure 4.4. *$(\lambda-\omega)$-system started with a shifted spiral, $G = SE(2)$.*

In polar coordinates this leads to the term

$$S(w,\gamma)\lambda = -\lambda_1 w_\varphi - \begin{pmatrix} w_r & \frac{1}{r} w_\varphi \end{pmatrix} \varrho_{-\varphi-\theta} \begin{pmatrix} \lambda_2 \\ \lambda_3 \end{pmatrix} - \varrho_{\frac{\pi}{2}} w \; \lambda_4.$$

Using the full symmetry group $SE(2) \times S^1$ immediately freezes the spiral wave in the shifted



(a) Frozen spiral

(b) Evolution of $\lambda(t)$

Figure 4.5. *$(\lambda - \omega)$-system started with a shifted spiral, $G = SE(2) \times S^1$. Clicking on the above image shows the associated movie that compares the evolution of the frozen system for both groups; the ϕ variable is indicated by a white bar.*

position. Also the velocities $\lambda_1 = \dot{\theta}$, $(\lambda_2, \lambda_3) = \dot{b}$, $\lambda_4 = \dot{\phi}$ stabilize at the stationary values $\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3 \approx 10^{-3}$, $\bar{\lambda}_4 > 0$ (see Figure 4.5) and there are no boundary effects as for $G = SE(2)$ above.

It seems surprising that the smaller group $G = SE(2)$ is sufficient to freeze this spiral wave, as shown in Figure 4.4. The reason is as follows: The relative equilibrium $u(x,t)$ has a special symmetry $u(\varrho_\theta x, t) = e^{i\theta}u(x,t)$ (i.e., the stabilizer is nontrivial). This makes it possible to transfer a rotation in the image of u to a rotation in the argument. The slight differences between Figures 4.4(a) and 4.5(a) seem to result from the fact that the coefficient λ_1 of the convective term w_φ in (4.10) is very small for the larger group, whereas it is of order one for the smaller group.

4.3.2. Quintic Ginzburg–Landau system.

The quintic Ginzburg–Landau system (QGL) given by

$$u_t = \left(\beta + \frac{i}{2}\right)\Delta u - \delta u + (\epsilon + i)|u|^2 u - (\mu + i\nu)|u|^4 u, \quad u(x,t) \in \mathbb{C},$$

possesses strongly localized solutions, so-called spinning solitons. These occur at parameter values $\beta = \delta = \frac{1}{2}$, $\epsilon = 2.5$, $\mu = 1$, $\nu = 0.1$; see [7], [8]. The symmetry group is again the four-dimensional group $SE(2) \times S^1$; cf. (4.15). We choose the following numerical parameters $R = 20$, $\Delta r = \frac{1}{6}$, $\Delta\varphi = \frac{\pi}{40}$, $\Delta t = 0.771 \cdot 10^{-4}$. We start at an initial profile that is obtained by shifting $u_0(r,\varphi) = 0.2\, e^{i\varphi} r e^{-(r/7)^2}$ slightly to the right on the x-axis. Then a localized vortex solution develops the real part, which is displayed in Figure 4.6.

We take this solution as initial data for the comparison of the rotating wave and its frozen

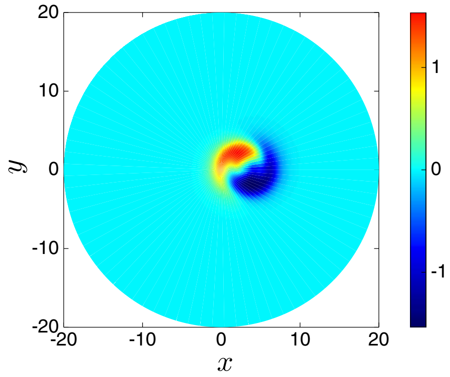
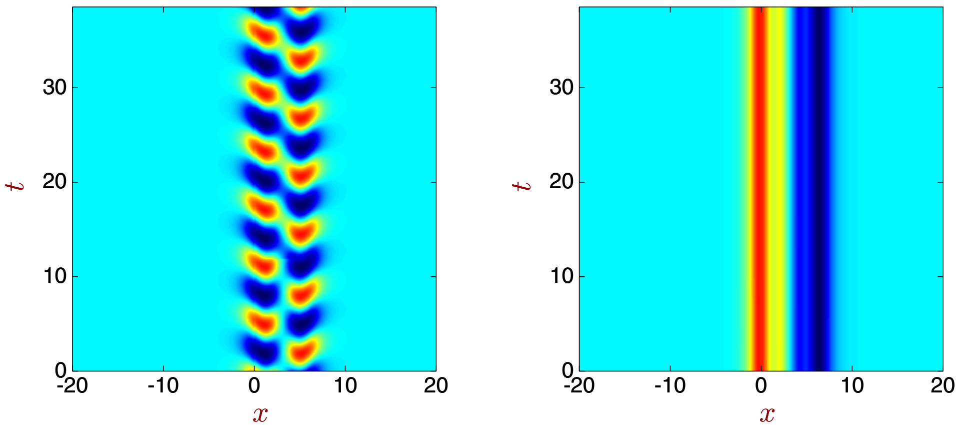


Figure 4.6. *QGL system.*



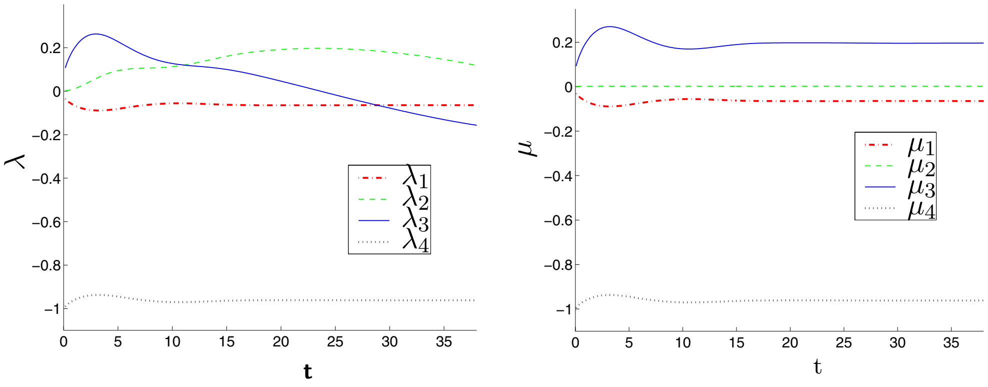
(a) Rotating spiral

(b) Frozen spiral

Figure 4.7. *Time evolution of a slice for the rotating and the frozen vortex for the QGL system. Clicking on the above image shows the associated movie of the original and the frozen system in the transformed variables (see Figure* 4.8(b)*).*

counterpart. Figure 4.7 shows the corresponding time evolution of a slice along the x-axis. The behavior of the λ parameters for the frozen system is displayed in Figure 4.8(a).

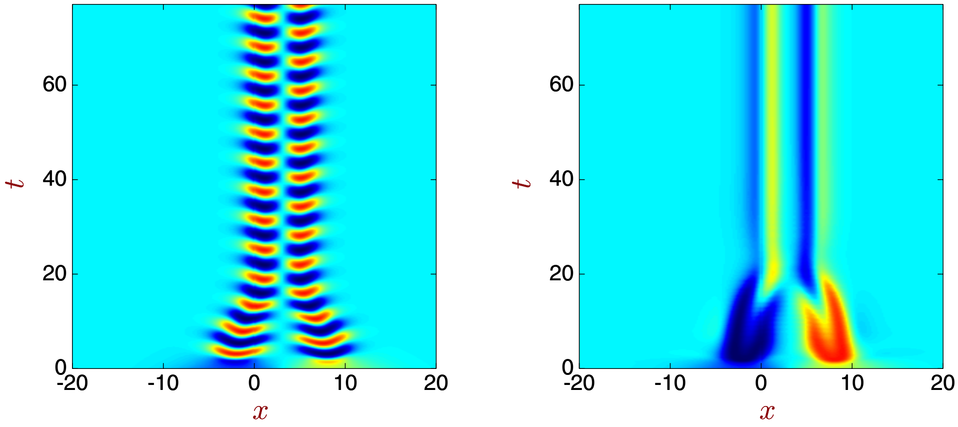
Their asymptotic behavior is better revealed by solving the reparametrized system (4.8), (4.9) (see Figure 4.8(b)). While the v-part is identical for both systems, the time plot of the parameters λ and μ shows a clear difference (see Figure 4.8). The values μ_2, μ_3 for the system (4.8), (4.9) stabilize after some time, showing that the center of rotation in (4.9) becomes constant. In contrast, the values λ_2, λ_3 for (4.6), (4.7) rotate according to $\binom{\lambda_2}{\lambda_3} = \varrho_\theta \binom{\mu_2}{\mu_3}$. The initial phase of this rotation is shown in Figure 4.8(a). The rotational velocities λ_1, μ_1 and λ_4, μ_4 are identical and rapidly stabilize at specific negative values.



(a) Nontransformed system

(b) Transformed system

Figure 4.8. *Frozen QGL system, evolution of parameters.*

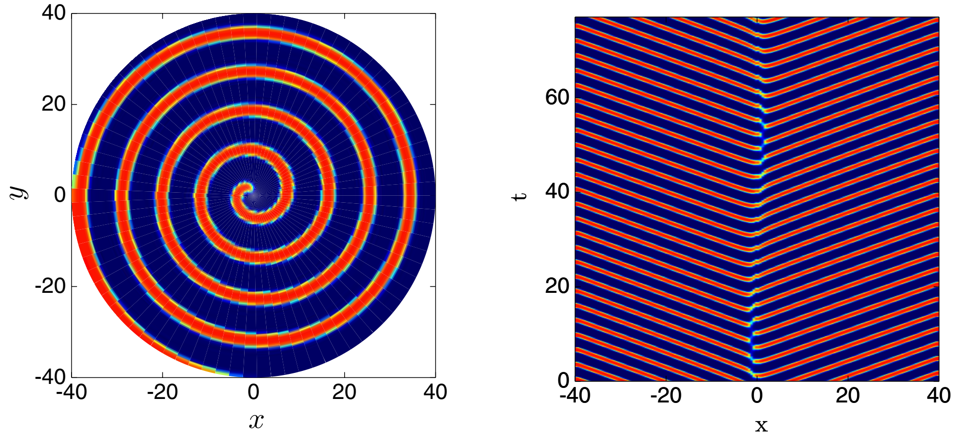


(a) Rotating spiral

(b) Frozen spiral

Figure 4.9. *Time evolution of a slice for the QGL system with initial data u_0 shifted to the right. Clicking on the above image shows the associated movie of the original and the frozen system in the transformed variables.*

Even for an initial value far away from any relative equilibrium, the longtime behavior of the frozen system and the nonfrozen system is similar. To this end we compare in Figure 4.9 the time evolution of both systems started with the shifted initial profile mentioned before. This initial function leads to a rotating vortex for the nonfrozen (see [7]) as well as for the frozen system.



(a) Spiral solution

(b) Time evolution of a slice

Figure 4.10. *Barkley system.*

4.3.3. Barkley's spiral system. Barkley's well-known system [3] is given by

$$u_t = \Delta u + \frac{1}{\epsilon} u(1-u)\left(u - \frac{v+b}{a}\right),$$
$$v_t = u - v.$$

The equation is equivariant with respect to the action of $SE(2)$; cf. (2.12), (4.4).

We test our method in the parameter regime of rigidly rotating waves with $\epsilon = \frac{1}{50}$ and $a = 0.75$, $b = 0.01$. The numerical parameters are $R = 40$, $\Delta r = 0.5$, $\Delta\varphi = \frac{\pi}{40}$, $\Delta t = 1.5421 \cdot 10^{-4}$.

In Figure 4.10(a) we show a snapshot of the u component of the spiral solution, and in Figure 4.10(b) the time evolution of a slice through u along the x-axis is displayed.

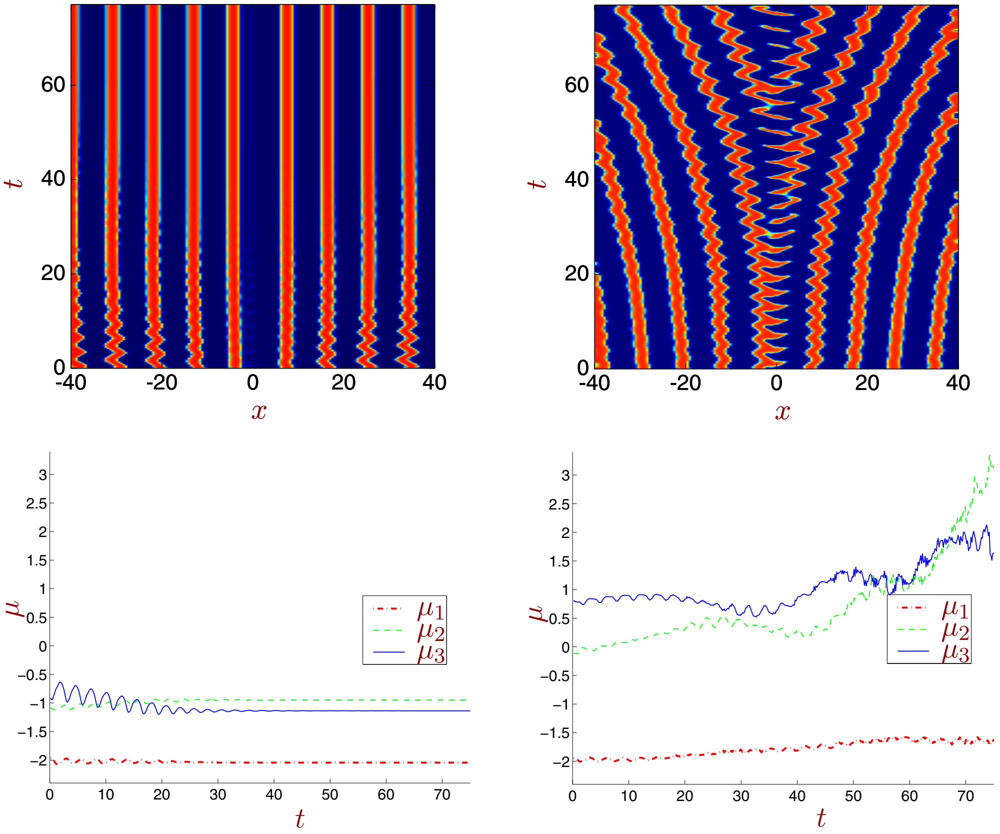
Since the system is of mixed hyperbolic-parabolic type, discretizing the convective terms with an upwind-downwind scheme becomes essential. In the previous examples, diffusion was strong enough to dominate convection introduced by the freezing procedure. For this example the contributions to first order derivatives are assembled in $b_1(r,\varphi)w_r$, resp., $b_2(r,\varphi)w_\varphi$, and approximated by

$$b_j(r,\varphi)D_\pm w = b_j(r,\varphi)[\chi(b_j(r,\varphi))D_+ w + (1-\chi(b_j(r,\varphi)))D_- w], \quad j = 1,2.$$

The symbols D_+, D_- denote forward and backward difference quotients in the r- or φ-direction, and the switching function χ is defined by

$$\chi(b) = (1 + \exp(-\beta\, b))^{-1}, \quad \beta = 0.2.$$

The value of β has been chosen in order to balance oscillations introduced by using centered diffences for the convective terms ($\beta = 0$) and artificial diffusion introduced by a strict switching rule (β large).



(a) L^2_η phase condition. The linked associated movie compares the rotating and the frozen system.

(b) L^2 phase condition. The linked associated movie compares the weighted and the nonweighted system.

Figure 4.11. *Frozen Barkley system, time evolution of a slice and of parameters* μ.

We start at a spiral wave, which rigidly rotates in the nonfrozen case. As Figure 4.10(b) shows, the spiral core still exhibits a slight oscillatory motion.

Figure 4.11(a) shows the result for the frozen system with a weighted L^2 norm, where the weight in (4.11) is $\eta(x) = e^{-0.5|x|}$. After some small initial oscillations, the wave eventually freezes and the parameters μ of the transformed system stabilize.

The importance of the weight in the phase condition is demonstrated in Figure 4.11(b), which shows the results for the L^2 inner product without weight. Now the spiral is perturbed by an oscillatory motion of the spiral core and finally drifts out of the region. The corresponding μ variables oscillate and drift away as well. These results suggest that the use of the weighted inner product for the phase condition keeps the continous spectrum in the left half-plane, while the L^2 inner product destabilizes the spectrum. Stabilization by the phase condition only seems to be different from shifting the whole spectrum to the left by considering the differential equation in a weighted function space as in [27]. The details of this mechanism need further investigation.

Initial values that are far away from relative equilibria lead to large differences of the time evolution for the frozen and the nonfrozen system. While the nonfrozen system develops a rigidly rotating spiral, we were not able to stabilize a corresponding frozen solution in a large time interval.

In summary, it turns out that freezing the spiral in Barkley's system is much more sensitive than in the previous examples. This is probably due to the mixed hyperbolic-parabolic type of the equation which requires more sophisticated numerical methods than our simple half-explicit scheme. Moreover, it seems quite a challenging task to freeze drifting spirals or recognize meandering spirals as periodic orbits.

Acknowledgments. The authors thank both referees for their useful suggestions which considerably helped in improving the first version of the paper.

REFERENCES

[1] D. Barkley, *A model for fast computer simulation of waves in excitable media*, Phys. D, 49 (1991), pp. 61–70.

[2] D. Barkley, *Linear stability analysis of rotating spiral waves in excitable media*, Phys. Rev. Lett., 68 (1992), pp. 2090–2093.

[3] D. Barkley, *Euclidean symmetry and the dynamics of rotating spiral waves*, Phys. Rev. Lett., 72 (1994), pp. 164–167.

[4] W.-J. Beyn, *The numerical computation of connecting orbits in dynamical systems*, IMA J. Numer. Anal., 10 (1990), pp. 379–405.

[5] V. N. Biktashev, A. V. Holden, and E. V. Nikolaev, *Spiral wave meander and symmetry of the plane*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 2433–2440.

[6] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, World Scientific, River Edge, NJ, 2000.

[7] L.-C. Crasovan, B. A. Malomed, and D. Mihalache, *Stable vortex solitons in the two-dimensional Ginzburg–Landau equation*, Phys. Rev. E, 63 (2001), 016605.

[8] L.-C. Crasovan, B. A. Malomed, and D. Mihalache, *Spinning solitons in cubic-quintic nonlinear media*, Pramana J. Phys., 57 (2001), pp. 1041–1059.

[9] M. Dellnitz, M. Golubitsky, A. Hohmann, and I. Stewart, *Spirals in scalar reaction diffusion equations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 5 (1995), pp. 1487–1501.

[10] B. Fiedler, B. Sandstede, A. Scheel, and C. Wulff, *Bifurcation from relative equilibria of noncompact group actions: Skew products, meanders, and drifts*, Doc. Math., 1 (1996), pp. 479–505.

[11] B. Fiedler and D. Turaev, *Normal forms, resonances, and meandering tip motions near relative equilibria of Euclidean group actions*, Arch. Ration. Mech. Anal., 145 (1998), pp. 129–159.

[12] M. Field, *Symmetry Breaking for Compact Lie Groups*, Mem. Amer. Math. Soc. 120, no. 574, AMS, New York, 1996.

[13] M. Golubitsky, V. G. LeBlanc, and I. Melbourne, *Meandering of the spiral tip: An alternative approach*, J. Nonlinear Sci., 7 (1997), pp. 557–586.

[14] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, SIAM, Philadelphia, 2000.

[15] E. Hairer, C. Lubich, and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods*, Lecture Notes in Math. 1409, Springer-Verlag, Berlin, 1989.

[16] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, 1996.

[17] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin, 1981.

[18] Y. Kuramoto and S. Koga, *Turbulized rotating chemical waves*, Progr. Theoret. Phys., 66 (1981), pp. 1081–1085.

[19] A. Mielke, *The Ginzburg–Landau equation in its role as a modulation equation*, in Handbook of Dynamical Systems, Vol. II, B. Fiedler, ed., North-Holland, Amsterdam, 2002, pp. 759–834.

[20] A. Mielke and G. Schneider, *Attractors for modulation equations on unbounded domains—existence and comparison*, Nonlinearity, 8 (1995), pp. 743–768.

[21] R. M. Miura, *Accurate computation of the stable solitary wave for the FitzHugh–Nagumo equations*, J. Math. Biol., 13 (1982), pp. 247–269.

[22] J. D. Murray, *Mathematical Biology*, Springer, Berlin, 1989.

[23] E. V. Nikolaev, V. N. Biktashev, and A. V. Holden, *On bifurcations of spiral waves in the plane*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 9 (1999), pp. 1501–1516.

[24] C. W. Rowley, I. G. Kevrekidis, J. E. Marsden, and K. Lust, *Reduction and reconstruction for self-similar dynamical systems*, Nonlinearity, 16 (2003), pp. 1257–1275.

[25] C. W. Rowley and J. E. Marsden, *Reconstruction equations and the Karhunen–Loève expansion for systems with symmetry*, Phys. D, 142 (2000), pp. 1–19.

[26] B. Sandstede, *Convergence estimates for the numerical approximation of homoclinic solutions*, IMA J. Numer. Anal., 17 (1997), pp. 437–462.

[27] B. Sandstede and A. Scheel, *Absolute versus convective instability of spiral waves*, Phys. Rev. E (3), 62 (2000), pp. 7708–7714.

[28] B. Sandstede, A. Scheel, and C. Wulff, *Dynamics of spiral waves on unbounded domains using center-manifold reductions*, J. Differential Equations, 141 (1997), pp. 122–149.

[29] B. Sandstede, A. Scheel, and C. Wulff, *Bifurcations and dynamics of spiral waves*, J. Nonlinear Sci., 9 (1999), pp. 439–478.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 2, pp. 117–160

A Rigorous Numerical Method for the Global Analysis of Infinite-Dimensional Discrete Dynamical Systems[*]

S. Day[†], O. Junge[‡], and K. Mischaikow[§]

Abstract. We present a numerical method to prove certain statements about the global dynamics of infinite-dimensional maps. The method combines set-oriented numerical tools for the computation of invariant sets and isolating neighborhoods, the Conley index theory, and analytic considerations. It not only allows for the detection of a certain dynamical behavior, but also for a precise computation of the corresponding invariant sets in phase space. As an example computation we show the existence of period points, connecting orbits, and chaotic dynamics in the Kot–Schaffer growth-dispersal model for plants.

Key words. Conley index, dynamical system, numercial method, infinite-dimensional

AMS subject classifications. 37B10, 37B30, 37B50, 37C25, 37C29, 37C70, 37L65, 37M99, 55U99, 65R20

DOI. 10.1137/030600210

1. Introduction.

The techniques described in this paper are motivated by the following three observations:

O1. Most of our knowledge concerning the global dynamics of specific nonlinear systems comes from numerical simulation and as such is lacking in mathematical rigor. In an attempt to rectify this, over the past decade a growing set of techniques has been developed that lead to computer assisted proofs of dynamical structures in low- (typically 2 or 3) dimensional systems (see [18, 24, 1] and references therein).

O2. Modeling of phenomena where spatial effects are essential leads to infinite-dimensional systems such as partial or functional differential equations, or infinite-dimensional maps. However, within these systems the dynamical structures of interest are often low-dimensional, e.g., fixed points, periodic orbits, homoclinic and heteroclinic orbits, horseshoes, or low-dimensional strange attractors.

O3. In dynamical systems the central objects of interest are invariant sets, i.e., collections of orbits which exist for all time. For a wide variety of infinite-dimensional systems, individual solutions which exist globally in time are more regular than the typical functions of the natural phase space (see [9] and references therein).

[*] Received by the editors May 9, 2003; accepted for publication (in revised form) by M. Golubitsky November 21, 2003; published electronically May 19, 2004. The first and third authors were partially supported by NSF grants DMS 0107396, DMS 9805584, and INT 9910027. The second author was supported by DFG grant JU 439/1-1 and DAAD/NSF grant D/0103717.

http://www.siam.org/journals/siads/3-2/60021.html

[†] Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, GA 30322 (sday.math03@gtalumni.org). Current address: Department of Mathematics, Cornell University, Ithaca, NY 14853.

[‡] Institute of Mathematics, University of Paderborn, 33095 Paderborn, Germany (junge@upb.de).

[§] Center for Dynamical Systems and Nonlinear Studies, Georgia Institute of Technology, Atlanta, GA 30322 (mischaik@math.gatech.edu).

Keeping these observations in mind, our goal is to provide a computationally cheap but accurate numerical method that can be used to prove existence theorems for specific infinite-dimensional maps. More precisely, the techniques that we describe are designed for continuous functions $\Phi : X \to X$, where X is a Hilbert space for which we have an explicit complete orthogonal basis $\{\varphi_k \mid k = 0, 1, 2, \ldots\}$ and the nonlinear terms are polynomial in nature. To provide a concrete demonstration of these ideas we will consider the Kot–Schaffer [15] growth-dispersal model for plants. This consists of a map $\Phi : L^2([-\pi, \pi]) \to L^2([-\pi, \pi])$ of the form

$$\Phi[a](y) := \frac{1}{2\pi} \int_{-\pi}^{\pi} b(x, y) g[a](x) dx, \tag{1.1}$$

with dispersal kernel $b(x, y) = b(x - y)$ and (polynomial) growth function g. In the following example, $g[a](x) := \mu\ a(x)(1 - \frac{a(x)}{c(x)})$ with $\mu > 0$ and $c \in L^2([-\pi, \pi])$. Observe that the regularity of this map is determined by the regularity of the dispersal kernel b and the spatial heterogeneity of c in the nonlinear term.

There are three obvious difficulties that need to be overcome to achieve our goal:

1. Because of the finite nature of a computer, it is impossible to compute directly on an infinite-dimensional system. Therefore, it is necessary to use an appropriate finite-dimensional reduction.
2. Given a finite-dimensional system, we need to be able to perform two tasks. The first is to locate the different dynamical objects. Since in many cases these objects are dynamically unstable, this is not a trivial task. The second is to rigorously verify that these dynamical structures exist for the finite-dimensional system.
3. We need to be able to lift the results of the finite-dimensional computations to the full infinite-dimensional system.

How these difficulties can be dealt with in a systematic and computationally efficient manner is the subject of this paper and as such is a natural extension of [26]. As the reader might expect, some of the details are fairly technical in nature, and therefore, we take the opportunity of this introduction to provide a broad outline of the procedures which will be developed in the following sections.

Let us begin with a reasonably abstract description of what will be done. We think of $\Phi : X \to X$ as generating a dynamical system with $a' = \Phi(a)$. Recall that $\{\varphi_k \mid k = 0, 1, 2, \ldots\}$ is a complete orthogonal basis for X. Let

$$P_m : X \to X_m := \operatorname{span}\{\varphi_k \mid k = 0, 1, \ldots, m - 1\}$$

be the orthogonal projection onto the first m modes. The standard Galerkin procedure suggests replacing the study of Φ by that of the map $f^{(m)} : X_m \to X_m$, where $f^{(m)} := P_m \circ \Phi$.

The problem is that if we study the dynamics using $f^{(m)}$, then we do not have any information concerning the errors introduced by the reduction to X_m and by the projection P_m. Observe that, to get around this problem, we can write

$$\Phi(a) = \Phi(P_m a) + (\Phi(a) - \Phi(P_m a)). \tag{1.2}$$

In general we cannot hope to determine the right-hand term exactly. However, if we restrict our attention to a "small" set of a, then we may be able to obtain a useful bound on this term. With this in mind, let W be a compact subset of X_m and let V be a compact subset of $(I - P_m)X$. Then,

$$Z := W \times V$$

is a compact subset of X.

Now assume that it can be shown that for all $a \in Z$,

$$\|\Phi(a) - \Phi(P_m a)\| < \varepsilon.$$

Then, for all $a \in Z$, $\Phi(a)$ lies within an ε-ball of $\Phi(P_m a)$. We want to recast these statements about bounds into the language of dynamical systems. Furthermore, we want this dynamics to be finite-dimensional so that we can effectively analyze it. This leads us to consider multivalued or set-valued maps $F : W \rightrightarrows \mathbb{R}^m$ with the property that for all $a \in Z$,

$$P_m \Phi(a) \in F(P_m a). \tag{1.3}$$

Perhaps it is worth noting at this point that if the images of F are "too large," then we will not be able to extract useful information from it. Thus, obtaining good bounds on $\|\Phi(a) - \Phi(P_m a)\|$ is essential.

At this point we have introduced two functions, the continuous map $f^{(m)} = P_m \circ \Phi : W \to \mathbb{R}^m$, which we do not know explicitly, and a multivalued map $F : W \rightrightarrows \mathbb{R}^m$, which encloses $f^{(m)}$ in the sense that $f^{(m)}(a) \in F(P_m a)$. It is the function F, which implicitly contains the error estimates, that we would like to analyze. However, to directly manipulate an object with the computer that object needs to have a combinatorial structure. With this in mind, W is decomposed into a cubical complex on which a combinatorial multivalued map $\mathcal{F}$ that takes grid elements to sets of grid elements is defined. Since each grid element corresponds to a set in W, it is easy to pass from the combinatorial map $\mathcal{F}$ to the multivalued map F.

This is perhaps a good point at which to emphasize the notational conventions adopted for this paper:

- Calligraphic characters represent combinatorial objects or maps.
- Capital letters refer to topological sets or set-valued maps.
- Single-valued maps (with the exception of Φ) are denoted by lowercase letters.

The discussion up to this point has described how one proceeds from the infinite-dimensional problem to a combinatorial object that can be analyzed using the computer. The question that remains is how to use this combinatorial information to draw conclusions about the dynamics of Φ. The key tool is the Conley index theory, which is a topological generalization of Morse theory. In particular, it can be expressed in terms of homology, which is a combinatorial algebraic topological theory. Furthermore, the index can be used to prove the existence of specific dynamical structures such as fixed points, periodic orbits, heteroclinic orbits, and shift dynamics.

As will be detailed later, $\mathcal{F}$ is used to construct isolating neighborhoods and index pairs, and finally to compute the associated Conley index for the map $f^{(m)}$. The important theoretical considerations are that one can pass from $\mathcal{F}$ to a multivalued map F, which is an enclosure of $f^{(m)}$, and that the Conley index information is preserved through this transition.

The final step is to show that there are conditions under which the Conley index of $f^{(m)}$ is equivalent to the Conley index of Φ restricted to Z. However, since Z is compact, the Conley index theory can be applied immediately to draw conclusions about the existence of dynamic structures for Φ.

At this level of abstraction, it is probably not even clear what are the issues involved in implementing this approach. With this in mind, we shall provide a broad outline of the procedures in the context of the Kot–Schaffer map (1.1).

1.1. Finite-dimensional reduction. We begin with the reduction to the finite-dimensional system, using Fourier modes to decompose $X = L^2$. In particular, letting $\varphi_k(x) := e^{ikx}$, (1.1) becomes equivalent to the countable system of maps,

$$a'_k = \mu b_k \left[a_k - \sum_{j+l+n=k} c_j a_l a_n \right], \quad k \in \mathbb{Z}, \tag{1.4}$$

where $a_k, b_k, c_k \in \mathbb{C}$ are the coefficients of the Fourier expansions of a, b, and c^{-1}, respectively. For simplicity, we restrict ourselves to the case $a_k = a_{-k}$, $b_k = b_{-k}$, and $c_k = c_{-k}$ for all $k \in \mathbb{Z}$. Therefore, (1.4) reduces to the system

$$a'_k = f_k(a) := \mu b_k \left[a_k - \sum_{j+l+n=k} c_{|j|} a_{|l|} a_{|n|} \right], \quad k = 0, 1, 2, \dots . \tag{1.5}$$

The resulting finite-dimensional system $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ upon which our numerical computations will be based is given by

$$a'_k = f_k^{(m)}(a_0, \dots, a_{m-1}) := \mu b_k \left[a_k - \sum_{\substack{j+l+n=k \\ 0 \le |l|, |n| \le m-1}} c_{|j|} a_{|l|} a_{|n|} \right], \tag{1.6}$$

where $k = 0, 1, \dots, m-1$.

Of course, in the end we will need to be able to justify that computation with (1.6) allows us to draw conclusions about the dynamics of (1.1). With this in mind, we rewrite the form of the maps as in (1.2) and define

$$f_k^{(m+)}(a) := f_k(a) - f_k^{(m)}(a_0, \dots, a_{m-1}).$$

Then, the full system becomes

$$a'_k = f_k^{(m)}(a_0, \dots, a_{m-1}) + f_k^{(m+)}(a), \qquad k = 0, 1, 2, \dots . \tag{1.7}$$

As was indicated earlier, we need to be able to bound values of the $f_k^{(m+)}$ term when we restrict our attention to specific compact subsets of L^2.

It is easy to check that the specific form of the nonlinear term in (1.6) comes from the fact that Φ has a quadratic nonlinearity. More generally, for a monomial of the form $c(x)a(x)^p$, the corresponding terms are

$$b_k \sum_{n_0,\dots,n_{p-1}\in\mathbb{Z}} c_{n_0} a_{n_1} \cdots a_{n_{p-1}} a_{k-(n_0+\cdots+n_{p-1})}.$$

In section 5 we will prove the following fundamental estimates (in slightly different forms).

Proposition 1.1. *Assume that there exist constants $s>1$, $C>0$, and $A>0$ such that*

$$|c_k| \le \frac{C}{|k|^s} \qquad \text{and} \qquad |a_k| \le \frac{A}{|k|^s}, \quad k \neq 0. \tag{1.8}$$

Then, for all $k \in \mathbb{Z}$,

$$\left| \sum_{n_0,\dots,n_{p-1}\in\mathbb{Z}} c_{n_0} a_{n_1} \cdots a_{n_{p-1}} a_{k-(n_0+\cdots+n_{p-1})} \right| \le \begin{cases} \frac{\alpha^p A^p C}{|k|^s} & k \neq 0, \\ \alpha^p A^p C & k = 0, \end{cases}$$

where $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$.

Proposition 1.2. *Assume that there exist constants $s>1$, $C>0$, and $A>0$ such that*

$$|c_k| \le \frac{C}{s^{|k|}} \qquad \text{and} \qquad |a_k| \le \frac{A_s}{s^{|k|}}. \tag{1.9}$$

Now fix $\gamma > 1$. Then, for all $k \in \mathbb{Z}$,

$$\left| \sum_{n_0,\dots,n_{p-1}\in\mathbb{Z}} c_{n_0} a_{n_1} \cdots a_{n_{p-1}} a_{k-(n_0+\cdots+n_{p-1})} \right| \le \frac{\alpha^p A^p C}{s^{|k|}} \gamma^{|k|},$$

where $\alpha = \frac{2}{\ln s} + \frac{\gamma}{\ln(\gamma)}$.

The hypotheses of these propositions may appear artificial until one recognizes that they are regularity conditions. In particular, the assumption on c_k is an explicit assumption on the regularity of the spatial heterogeneity of the nonlinear term in (1.1). The assumption on the a_k, however, needs greater justification. Obviously, if a is a typical element of L^2, then neither (1.8) nor (1.9) will be satisfied. However, our interest lies in elements which belong to invariant sets. As was mentioned in observation O3, such elements often possess considerable regularity. This property will be shown explicitly in Lemma 5.1 for the Kot–Schaffer map.

To make use of Proposition 1.1 or 1.2 we need to determine the constants A, s, and m. As is explained in detail in section 3, this is done via simulations of the finite-dimensional system $f^{(L)}: \mathbb{R}^L \to \mathbb{R}^L$, where L is large. In particular, we choose an initial condition $a^0 = (a^0_0, \dots, a^0_{L-1})$ and examine the iterates $a^1 := f^{(L)}(a^0), a^2 := f^{(L)}(a^1), a^3 := f^{(L)}(a^2), \dots$. We can then divide the individual coordinates into two groups as follows. First, we look for those coordinates that fluctuate on a predetermined scale under the iterations. Typically, these involve the "lower modes"; i.e., there exists an integer m such that the only a_k's which change significantly under iterations of the map are those for which $k < m$. Then, one examines the remaining a^n_k's for $m \le k < L$ and uses this information to choose the constants A_s and s.

Observe that at this step we are not making any claims of rigor. This is an important point. Experience suggests that rigorous computations are typically expensive. Therefore, our approach is to use simulations as much as possible and at the final step to perform a rigorous verification of the results.

Using the above procedure we arrive at a finite-dimensional system $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$. As was indicated earlier, the first problem we need to consider is how to find the dynamical structures of interest. From the simulation, we choose a range of values of the coordinates a_k, $k = 0, \ldots, m-1$; i.e., we choose constants a_k^- and a_k^+ such that we are only interested in $a_k \in \widetilde{a}_k := [a_k^-, a_k^+]$. Let

$$W := \prod_{k=0}^{m-1} [a_k^-, a_k^+]$$

and consider $f^{(m)} : W \to \mathbb{R}^m$.

Having fixed W, we can now determine the constant A in Proposition 1.1 or 1.2. Observe that by this procedure we have effectively restricted our attention to the dynamics of Φ on the set

$$Z := W \times \prod_{k=m}^{\infty} \widetilde{a}_k, \tag{1.10}$$

where $\widetilde{a}_k := [-\frac{A_s}{|k|^s}, \frac{A_s}{|k|^s}]$ or $\widetilde{a}_k := [-\frac{A_s}{s^{|k|}}, \frac{A_s}{s^{|k|}}]$ depending on whether one assumes polynomial or exponential decay rates. Clearly, for s sufficiently large, Z is a compact set. The obvious question at this point is whether this assumption restricts the types of invariant sets that one can capture. Here again observation O3 comes into play. For a large variety of systems, elements of invariant sets possess significant regularity properties, and therefore, will lie in a set of the form of Z.

We can then use Proposition 1.1 or 1.2 to determine bounds $\varepsilon_k^{(m+)} > 0$ for $f_k^{(m+)}$, $k = 0, 1, 2, \ldots$, i.e., for all $a \in Z$, $|f_k^{(m+)}(a)| < \varepsilon_k^{(m+)}$.

1.2. Analysis of finite-dimensional systems.

Up to this point in the analysis, most of what has been described is easily subsumed within the standard Galerkin approximation techniques. We now change the perspective in our analysis of the finite-dimensional system. While we have an explicit representation of $f_k^{(m)}$, we are really interested in the dynamics of the system generated by the f_k. Unfortunately, the best we have been able to do is determine a bound $\varepsilon_k^{(m+)}$ on $f_k^{(m+)}$. To incorporate the errors into our analysis, we adopt the philosophy that our finite-dimensional system should be viewed as a set-valued map satisfying (1.3).

Observe that, even if the multivalued map satisfies (1.3), it is probably not directly amenable to analysis by the computer since it does not have a combinatorial structure. To resolve this problem we partition W into a *cubical grid*,

$$\mathcal{G}^{(\sigma)} := \left\{ G = \prod_{k=0}^{m-1} a_k^- + \left[\frac{i_k}{2^\sigma}, \frac{i_k+1}{2^\sigma}\right] w_k \;\middle|\; i_k \in \{0, \ldots, 2^\sigma - 1\} \right\},$$

where $w_k = a_k^+ - a_k^-$ and $\sigma \in \mathbb{N}$. Given $\mathcal{B} \subset \mathcal{G}^{(\sigma)}$, let $|\mathcal{B}|$ denote the union of all the elements in $\mathcal{B}$ viewed as subsets of $\mathbb{R}^m$. We will use a multivalued map $\mathcal{F}^{(\sigma)} : \mathcal{G}^{(\sigma)} \rightrightarrows \mathcal{G}^{(\sigma)}$ that satisfies the following property. For every $G \in \mathcal{G}^{(\sigma)}$,

$$\left\{ (f_0, \dots, f_{m-1})(a) \;\middle|\; a \in G \times \prod_{k=m}^{\infty} \widetilde{a}_k \right\} \subset \operatorname{int}(|\mathcal{F}^{(\sigma)}(G)|). \tag{1.11}$$

Observe that this is just a restatement of (1.3).

For the moment, we will treat σ as fixed and so simplify the notation by setting $\mathcal{F} = \mathcal{F}^{(\sigma)}$ and $\mathcal{G} = \mathcal{G}^{(\sigma)}$. Since $\mathcal{F}$ is defined on the set of cubes in $\mathcal{G}$ and has images that consist of sets of these cubes, it is a combinatorial object. Hence it can be analyzed directly by the computer. A precise description of the construction of $\mathcal{F}$ is given in section 3. For now it is sufficient to observe that it is determined by the bounds $\varepsilon_k^{(m+)}$ on $f_k^{(m+)}$ and estimates on $f_k^{(m)}$ for $k = 0, \dots, m-1$. Furthermore, the estimates on $f_k^{(m)}$ depend on the size of the elements of $\mathcal{G}$. In particular, larger σ leads to better estimates.

$\mathcal{F}$ is the finite-dimensional dynamical system whose structure we will study. Since the reader may not have encountered multivalued dynamical systems before, we include a few essential definitions at this point. A *full combinatorial trajectory of $\mathcal{F}$ through $G \in \mathcal{G}$* is a bi-infinite sequence $\gamma_G : \mathbb{Z} \to \mathcal{G}$ satisfying $\gamma_G(0) = G$ and $\gamma_G(n+1) \in \mathcal{F}(\gamma_G(n))$ for all $n \in \mathbb{Z}$. $\mathcal{S} \subset \mathcal{G}$ is a *combinatorial invariant set* if for every $G \in \mathcal{S}$ there exists a full solution $\gamma_G : \mathbb{Z} \to \mathcal{S}$.

Even though $\mathcal{F}$ is a finite combinatorial object, and therefore, capable of being examined directly by the computer, the associated complexity is impractical. Observe that the number of elements of $\mathcal{G}$ is $(2^\sigma)^{\dim W}$. To avoid this problem we make use of a subdivision algorithm as described in [4] to find the maximal invariant set $\mathcal{A}$ of $\mathcal{F}$ in $\mathcal{G}$. The computational effort to find $\mathcal{A}$ is of the same order as the number of elements in $\mathcal{A}$, which is approximately $(2^\sigma)^{\dim \mathcal{A}}$. This is the point at which the observation O2 comes into play. If we are looking for low-dimensional dynamical structures, then we can hope to do so in a computationally efficient manner even if m, the dimension of the approximation, is large.

The condition (1.11) guarantees that $\mathcal{F}$ acts as an outer approximation of the map $f^{(m)}$. However, knowledge of trajectories of $\mathcal{F}$ does not directly lead to information about the existence of trajectories of $f^{(m)}$. For example, the existence of a fixed point for $\mathcal{F}$, i.e., a full solution $\gamma_G : \mathbb{Z} \to \mathcal{G}$ that has the form $\gamma_G(n) = G$ for all $n \in \mathbb{Z}$, does not imply that $f^{(m)}$ possesses a fixed point. To obtain this information requires the use of algebraic topology and, in particular, the Conley index theory.

The starting point for the computational version of this theory is the notion of a *combinatorial isolating neighborhood*, a finite set whose maximal invariant set lies strictly in its interior. To make this precise in the setting of a cubical grid, given $\mathcal{B} \subset \mathcal{G}$, let

$$o(\mathcal{B}) := \{ G \in \mathcal{G} | \, |G| \cap |\mathcal{B}| \neq \emptyset \}.$$

Observe that $o(\mathcal{B})$ is the smallest neighborhood of $\mathcal{B}$ that can be represented using elements of the cubical grid. Given $\mathcal{I} \subset \mathcal{G}$, the *maximal combinatorial invariant set* in $\mathcal{I}$ is

$$\operatorname{Inv}(\mathcal{I}, \mathcal{F}) := \{ G \in \mathcal{I} \mid \text{there exists a full trajectory } \gamma_G : \mathbb{Z} \to \mathcal{I} \}.$$

$\mathcal{I}$ is a *combinatorial isolating neighborhood* if

$$o(\operatorname{Inv}(\mathcal{I})) \subset \mathcal{I}. \tag{1.12}$$

An important result following from (1.11) is that if $\mathcal{I}$ is an isolating neighborhood for $\mathcal{F}$, then $|\mathcal{I}|$ is an isolating neighborhood for $f^{(m)}$.

To each isolating neighborhood I of $f^{(m)}$ one can assign a Conley index, which is denoted by $\chi_*(I, f^{(m)})$. To compute this index one needs to first construct an index pair. As was indicated earlier, these computations are done using the combinatorial multivalued map for which we make use of the following definition.

Definition 1.3. Let $\mathcal{I}$ be an isolating neighborhood for $\mathcal{F}$. A pair $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_0)$ of subsets $\mathcal{N}_0 \subset \mathcal{N}_1 \subset \mathcal{I}$ is a *combinatorial index pair* if the following conditions are satisfied:

1. $\mathcal{F}(\mathcal{N}_i) \cap \mathcal{I} \subset \mathcal{N}_i$;
2. $\mathcal{F}(\mathcal{N}_1 \setminus \mathcal{N}_0) \subset \mathcal{I}$;
3. $o\left(\operatorname{Inv}(\mathcal{I}, \mathcal{F})\right) \subset \mathcal{N}_1 \setminus \mathcal{N}_0$.

Of course our immediate interest does not lie with the multivalued map $\mathcal{F}$, but rather with the single-valued map $f^{(m)}$. We will make use of the fact that an index pair for $\mathcal{F}$ is also an index pair for $f^{(m)}$. In fact, Szycmzak [22, 23] has shown that we can use a slightly weaker construction based on $\mathcal{F}$. Let $\mathcal{S}$ be an isolated invariant set for $\mathcal{F}$. Define,

$$\mathcal{N}_1 := \mathcal{S} \cup \mathcal{F}(\mathcal{S}) \qquad \text{and} \qquad \mathcal{N}_0 := \mathcal{N}_1 \setminus \mathcal{S}. \tag{1.13}$$

Then, $|\mathcal{N}| = (|\mathcal{N}_1|, |\mathcal{N}_0|)$ is an index pair for $f^{(m)}$.

Further details concerning this index are provided in section 2. For the moment, two remarks suffice. The first is that $\chi_*(|\mathcal{N}_1| \setminus |\mathcal{N}_0|, f^{(m)})$, the Conley index under $f^{(m)}$, can be computed using the combinatorial information of $\mathcal{F}$. The second is that the Conley index provides information about the structure of the associated maximal invariant set.

These remarks should make it clear that, in order to use the Conley index to understand the dynamics of $f^{(m)}$, it is essential that we are able to efficiently find isolating neighborhoods $\mathcal{I}$ of $\mathcal{F}$. Furthermore, we need to be able to find isolating neighborhoods that isolate specific dynamical objects of interest such as fixed points, periodic orbits, heteroclinic orbits, etc. Up to this point we have treated $\mathcal{F}$ as a dynamical system. However, to efficiently find specified orbits, it is useful to view $\mathcal{F}$ as a directed graph; the vertices correspond to the elements of $\mathcal{G}$, and if $G_1 \in \mathcal{F}(G_0)$, then there is a directed edge from G_0 to G_1. Full trajectories of $\mathcal{F}$ now correspond to infinite paths in the directed graph. These paths, or sets of paths, representing either recurrent sets or connecting orbits between recurrent sets, can be found using standard graph-theoretic algorithms (see [7] and references therein). In this paper we make use of the implementation in the software package GAIO as detailed in [5].

Observe that the above-mentioned paths correspond to a set of elements $\mathcal{I} \subset \mathcal{G}$. $\mathcal{I}$ is clearly an invariant set for $\mathcal{F}$, and hence, cannot be an isolating neighborhood. However, $o(\mathcal{I})$ is a candidate for being an isolating neighborhood. In particular, we can compute $\operatorname{Inv}(o(\mathcal{I}), \mathcal{F})$. If $o(\operatorname{Inv}(o(\mathcal{I}), \mathcal{F})) \subset \mathcal{I}$, then $\mathcal{I}$ is an isolating neighborhood. If not, then we can repeat the procedure starting with the larger invariant set $\operatorname{Inv}(o(\mathcal{I}), \mathcal{F})$. The specific algorithm is given in section 4.

1.3. Lifting to the full system. At this point we have described the finite-dimensional reduction and the computer-assisted analysis of the resulting finite-dimensional system. In particular, assume that we have found a pair of subsets $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_0)$ of $\mathcal{G}$ satisfying (1.13). As was mentioned above, this allows us to determine the Conley index $\chi_*(|\mathcal{N}_1| \setminus |\mathcal{N}_0|, f^{(m)})$. This is the information that we wish to lift to the full infinite-dimensional system.

Recall that we have restricted our attention to the dynamics in

$$Z = W \times \prod_{k=m}^{\infty} \widetilde{a}_k.$$

Let us further assume that for high modes, i.e., $k \geq m$, the system is contracting. More precisely, assume that for all $a \in Z$,

$$f_k(a) \subset \operatorname{int}(\widetilde{a}_k). \tag{1.14}$$

In practice this is verified using Proposition 1.1 or 1.2.

In this case it is easy to show (see Theorem 2.3) that

$$\widetilde{N} = (\widetilde{N}_1, \widetilde{N}_0) := \left(|\mathcal{N}_1| \times \prod_{k=m}^{\infty} \widetilde{a}_k, |\mathcal{N}_0| \times \prod_{k=m}^{\infty} \widetilde{a}_k \right)$$

is an index pair for Φ and that

$$\chi_*(\widetilde{N}_1 \setminus \widetilde{N}_0, \Phi) \cong \chi_*(|\mathcal{N}_1| \setminus |\mathcal{N}_0|, f^{(m)}).$$

Observe that both $\widetilde{N}_1$ and $\widetilde{N}_0$ are compact sets. Therefore, we can apply the classical results from the Conley index theory to draw conclusions about the dynamics of Φ.

1.4. Refinement of the combinatorial invariant set. We mentioned above that, given an isolating neighborhood, we have algorithms that compute the associated Conley index. As will become clear when we present the results from specific computations in section 6, this is by far the most computationally intensive step. Therefore, our strategy is to compute the Conley index using a minimal number of modes, i.e., as small an m as possible, and the fewest boxes, i.e., choosing σ as small as possible.

As was described earlier, we use the Conley index of an isolating neighborhood I to obtain information about $\operatorname{Inv}(I, f^{(m)})$. Since $\operatorname{Inv}(I, f^{(m)}) \subset I$, the diameters of the components of I provide us with bounds on the individual trajectories in $\operatorname{Inv}(I, f^{(m)})$. One of the other important properties of the Conley index is that it is an index of invariant sets; i.e., if I and I' are different isolating neighborhoods such that

$$\operatorname{Inv}(I, f^{(m)}) = \operatorname{Inv}(I', f^{(m)}),$$

then $\chi_*(I, f^{(m)}) = \chi_*(I', f^{(m)})$. This allows us to efficiently improve the bounds on the invariant sets as follows. Let $\mathcal{I} \subset \mathcal{G}^{(\sigma)}$ be an isolating neighborhood for the map $\mathcal{F}^{(\sigma)}$. Now choose $\sigma' > \sigma$. Then, there exists a set $\mathcal{B} \subset \mathcal{G}^{(\sigma')}$ such that $|\mathcal{B}| = |\mathcal{I}|$. Let $\mathcal{I}' := \operatorname{Inv}(\mathcal{B}, \mathcal{F}^{(\sigma')})$. Then, in general $|\mathcal{I}'| \subset |\mathcal{I}|$. However, $\chi_*(|\mathcal{I}'|) = \chi_*(|\mathcal{I}|)$. Thus, we retain the information about the invariant set, but with better bounds.

Still, there is a natural limit to how big we can choose σ' to be for a given m. This limit is essentially determined by the bounds $\varepsilon_k^{(m+)}$ on the terms $f_k^{(m+)}$ in (1.7)—note that one would not expect $|\mathcal{I}'|$ to be any smaller than $|\mathcal{I}|$ if the bounds $\varepsilon_k^{(m+)}$ are of the same order of magnitude as the size of the grid elements in $G^{(\sigma')}$. There are two ways to escape from this dilemma:

1. Improve the values a_k^-, a_k^+, $k \geq m$, i.e., make the intervals $[a_k^-, a_k^+]$ as small as possible—since this will decrease the bounds $\varepsilon_k^{(m)}$ via Propositions 1.1 and 1.2. Once a tighter neighborhood $\mathcal{I}'$ has been constructed, the intervals can be updated such that (1.14) is still satisfied.
2. Increase the number of modes m used in the computations—this will also decrease the bounds $\varepsilon_k^{(m+)}$. Again, we make use of the fact that by suitably lifting the isolating neighborhood $\mathcal{I}$ from $\mathbb{R}^m$ to $\mathbb{R}^{m+1}$, we retain the index information and thus the information about the invariant set. One may think that the computational effort increases dramatically when going to higher-dimensional phase spaces. This is not the case, as will be detailed in later sections and demonstrated in the examples section. In fact, due to the hierachical design of the refinement process and the way the combinatorial map $\mathcal{F}$ is computed, the computational effort grows only linearly in m.

1.5. Results.

In section 6 we are going to prove a couple of prototype results about the dynamics of the map Φ for different parameter values. These results are meant to serve as examples of what kind of statements may be obtained by our method and are, of course, by no means a complete description of the dynamics of Φ. We are going to sketch these results in the following; for the exact technical statements we refer to section 6.

Result 1. *For certain parameter values the map Φ possesses a fixed point a_1 and a period two point a_2, as well as an orbit limiting in backward time to a neighborhood of a_1 and in forward time to a neighborhood of a_2. We localize these objects up to an error of* 10^{-12}.

Result 2. *For the same parameter values, the map Φ possesses an invariant set on which it exhibits complicated dynamics. The topological entropy of Φ is at least* 1.2.

Result 3. *For certain parameter values, the map Φ possesses an invariant set on which Φ is semiconjugate to the subshift dynamics given by the transition graph T shown in Figure 1.1.*

2. The Conley index.

The Conley index theory is a key component in our approach to the rigorous analysis of high-dimensional dynamical systems. It provides the tool by which we can pass from combinatorial data to dynamical structures. We present it here (for the most part) as a black box; that is, notation, definitions, and key theorems are stated, but no attempt is made to motivate the underlying justification. In part, this is done for reasons of space but also to convey the fact that the computational methods being developed are sufficiently self-contained to be performed without knowledge of the index theory, and similarly, the index theory can be applied without reference to the details of the computations. As such, depending on the reader's preferences, this section can be skipped and only referred to for notation or critical results.

There are two issues that, in the context of our approach, are of particular importance. The first is robustness of the index with respect to perturbation. There are essentially two

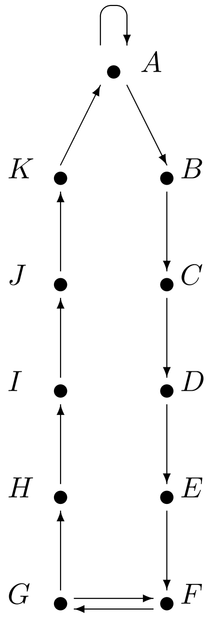


Figure 1.1. *The graph T.*

perturbations that need to be controlled: the numerical approximation used to study the finite-dimensional map, and the finite-dimensional approximation of the true infinite-dimensional dynamics. The second issue is that of using the index to recover information about the dynamics.

2.1. Robustness. As was indicated in the introduction, depending on our immediate needs, we use the original system to construct three types of dynamical systems in this paper: continuous maps, $f : Y \to Y'$; multivalued maps, $F : Y \rightrightarrows Y'$; and combinatorial maps, $\mathcal{F} : \mathcal{G} \rightrightarrows \mathcal{G}$. It is assumed that Y is a compact subset of Y' and that $\mathcal{G}$ is a cubical complex such that $|\mathcal{G}| = Y$.

Of course, the relationship between these maps is crucial. Throughout our discussion we require that (1.11) be satisfied. Observe that the process of passing from a combinatorial map to a multivalued map is quite simple. Given $\mathcal{F}$, define $F : |\mathcal{G}| \rightrightarrows |\mathcal{G}|$ by

$$F(y) := \bigcup_{y \in G} |\mathcal{F}(G)|. \tag{2.1}$$

Now, observe that a single-valued map $f : Y \to Y$ is a special case of a multivalued map $F : Y \rightrightarrows Y$. Therefore, any definition given in terms of a multivalued map is immediately applicable to continuous maps. In the introduction, we gave some of the fundamental definitions in the setting of combinatorial dynamics. We repeat them now in the context of multivalued maps.

A *full trajectory* of F is a bi-infinite sequence $\sigma_y : \mathbb{Z} \to Y$ satisfying $\sigma_y(0) = y$ and $\sigma_y(n+1) \in F(\sigma_y(n))$ for all $n \in \mathbb{Z}$. $S \subset Y$ is an *invariant set* if for every $y \in S$ there exists a full solution $\sigma_y : \mathbb{Z} \to S$. A compact set $I \subset Y$ is an *isolating neighborhood* if its maximal invariant set is contained in its interior, i.e.,

$$\operatorname{Inv}(I, F) := \{y \in I \mid \exists \sigma_y : \mathbb{Z} \to I\} \subset \operatorname{int}(I).$$

An *isolated invariant set* is the maximal invariant set of an isolating neighborhood.

The following definition is due to Szymczak [23].

Definition 2.1. A pair $N = (N_1, N_0)$ of compact sets is an *index pair* for F if the following conditions are satisfied:

1. $F(N_0 \cap \operatorname{cl}(N_1 \setminus N_0)) \cap N_1 \subset N_0$;
2. $F(N_1 \setminus N_0) \subset N_1$;
3. $I := \operatorname{cl}(N_1 \setminus N_0)$ is an isolating neighborhood under F.

Notice that if one applies the same idea used to transform the combinatorial map $\mathcal{F}$ to the multivalued map F as in (2.1), then the definitions for the combinatorial objects given in the introduction are consistent with these new definitions. In this way we can use graph-theoretic techniques to compute topological objects.

Let $F, F' : Y \rightrightarrows Y'$. F *encloses* F' if

$$F'(y) \subset F(y)$$

for every $y \in Y$. In the special case where F' is a continuous single-valued map, F' is called a *continuous selector* for F.

The following result is trivial but is of special interest in the case where F' is a continuous selector because it allows us to transfer the topological constructions for multivalued maps (which encode the errors) to continuous maps, which are not explicitly known.

Proposition 2.2. *Assume F encloses F'.*

1. *If σ_y is a trajectory for F', then σ_y is a trajectory for F.*
2. *If I is an isolating neighborhood for F, then I is an isolating neighborhood for F'.*
3. *If $N = (N_1, N_0)$ is an index pair for F, then N is an index pair for F'.*

Now consider the case of a continuous map $f : Y \to Y'$. Let $N = (N_1, N_0)$ be an index pair under f with isolating neighborhood $I := \operatorname{cl}(N_1 \setminus N_0)$. Let N_1/N_0 denote the quotient space obtained by identifying the elements of N_0 to a single point. Then, by [23, Theorem 1.4] f induces a continuous map

$$f_N : (N_1/N_0, [N_0]) \to (N_1/N_0, [N_0]).$$

On the level of homology, f_N induces a map

$$\chi_*(I, f) : H_*(N_1, N_0) \to H_*(N_1, N_0)$$

which, up to an equivalence class, is the Conley index of I. For details the reader is referred to [23, 8, 16].

A fundamental property of the Conley index which we will exploit heavily is the following. Let I be an isolating neighborhood for a multivalued map F taking contractible values. If g and f are continuous selectors for F, then their Conley indices are the same, i.e.,

$$\chi_*(I, f) \cong \chi_*(I, g).$$

Since F encodes the numerical errors, this is a statement about the robustness of the index with respect to numerical error.

We now have the following sequence of implications. Let $\mathcal{F}$ be a combinatorial map. As will be indicated in section 3, there are efficient algorithms to produce an isolating neighborhood $\mathcal{I}$ and index pair $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_0)$ for $\mathcal{F}$. Using (2.1) we obtain a multivalued map F for which $I = |\mathcal{I}|$ is an isolating neighborhood and $N = (N_1, N_0) = (|\mathcal{N}_1|, |\mathcal{N}_0|)$ is an index pair. Let f be a continuous selector for F. Then, I is an isolating neighborhood for f with an index pair $N = (N_1, N_0)$. An important remark is that one can use the combinatorial information of $\mathcal{F}$ and $\mathcal{N}$ to compute $H_*(N_1, N_0)$ and $\chi_*(I, f)$; see [17, 11].

We now focus our attention on a specific setting of relevance to this paper. Let X be a Hilbert space with an orthogonal basis (φ_k). Let $P_m : X \to X_m := \operatorname{span}\{\varphi_k \mid k = 0, \dots, m\}$ be a projection and let $Q_m = I - P_m$ be the projection onto the complementary subspace X'_m.

Let $W \subset X_m$ and $V \subset X'_m$ be compact subsets. In addition, assume that V is contractible. Let ∂V denote the boundary of $V \subset X'_m$ and let $Z = W \times V$.

Theorem 2.3. *Let $F' : Z \rightrightarrows X_m \times V$ and $F : W \rightrightarrows X_m$ be multivalued maps such that*

$$P_m F'(z) \subset F(P_m z) \qquad \forall\, z \in Z. \tag{2.2}$$

Assume that

$$Q_m F'(Z) \subset \operatorname{int}(V). \tag{2.3}$$

Then the following results hold:

(a) *If I is an isolating neighborhood for F, then $\hat{I} := I \times V$ is an isolating neighborhood for F'.*

(b) *If $N = (N_1, N_0)$ is an index pair for F, then $\hat{N} := (N_1 \times V, N_0 \times V)$ is an index pair for F'.*

(c) *Let g and g' be continuous selectors for F and F', respectively; then*

$$\chi_*(\hat{I}, g') \cong \chi_*(I, g).$$

Proof. (a) To show that $\hat{I}$ is an isolating neighborhood it is sufficient to show that $\partial(\hat{I}) \cap \operatorname{Inv}(\hat{I}, F') = \emptyset$. Consider $z \in \partial(\hat{I})$. Suppose first that $P_m z \in \partial I$ and let $\sigma_z : \mathbb{Z} \to X$ be a full trajectory through z under F', i.e., $\sigma_z(n+1) \in F'(\sigma_z(n))$ and $\sigma_z(0) = z$. Observe that, by (2.2), $P_m \sigma_z$ defines a full trajectory for F. However, $P_m z \in \partial I$. Since I is an isolating neighborhood for F, there exists $n \in \mathbb{Z}$ such that $P_m \sigma_z(n) \notin I$, and hence $\sigma_z(n) \notin \hat{I}$. If, however, $P_m z \notin \partial I$, then $Q_m z \in \partial V$. In this second case, (2.3) implies that $z \notin \operatorname{Inv}(\hat{I}, F')$.

(b) Suppose $N = (N_1, N_0)$ is an index pair for F. We now prove that $\hat{N} := (\hat{N}_1, \hat{N}_0)$, where $\hat{N}_i := N_i \times V$, is an index pair for F'. By Definition 2.1-3, $I := \operatorname{cl}(N_1 \setminus N_0)$ is an isolating neighborhood under F. By (a), $\hat{I} := \operatorname{cl}(\hat{N}_1 \setminus \hat{N}_0)$ is an isolating neighborhood for F' and condition (3) is satisfied for the pair $\hat{N}$. We now need to check that the remaining two conditions of Definition 2.1 are satisfied by $\hat{N} = (\hat{N}_1, \hat{N}_0)$.

(1) Let $z \in \hat{N}_0 \cap \operatorname{cl}(\hat{N}_1 \setminus \hat{N}_0)$. Then $P_m z \in N_0 \cap \operatorname{cl}(N_1 \setminus N_0)$ and using (2.2),

$$P_m \left(F'(z) \cap \hat{N}_1 \right) \subset F(P_m z) \cap N_1 \subset N_0 = P_m(\hat{N}_0).$$

By (2.3),

$$Q_m \left(F'(z) \cap \hat{N}_1 \right) = Q_m \left(F'(z) \right) \cap V \subset Q_m(\hat{N}_0).$$

Therefore, $F'(\hat{N}_0 \cap \mathrm{cl}\,(\hat{N}_1 \setminus \hat{N}_0)) \cap \hat{N}_1 \subset \hat{N}_0$.

(2) Let $z \in \hat{N}_1 \setminus \hat{N}_0$. By (2.2), $P_m F'(z) \subset F(P_m z)$. Since $P_m z \in N_1 \setminus N_0$, $F(P_m z) \subset N_1 = P_m \hat{N}_1$. Hence, $P_m F'(z) \subset P_m \hat{N}_1$. In addition, by (2.3) $Q_m F' z \subset \mathrm{int}\, V \subset Q_m \hat{N}_1$. Therefore, $F'(\hat{N}_1 \setminus \hat{N}_0) \subset \hat{N}_1$.

(c) By assumption, V is contractible; i.e., there exists a continuous map $h : V \times [0,1] \to V$ such that $h(\cdot, 0) = I_V$, the identity on V, and $h(\cdot, 1) = v_0 \in V$. Define $H : Z \times [0,1] \to Z$ by $H(z,s) = P_m g'(z) + h(Q_m g'(z), s)$. Observe that $H(z,0) = g'(z)$ and $H(\cdot, 1) : Z \to W \times \{v_0\}$. Furthermore,

$$P_m H(z,s) \in F(P_m z) \qquad \forall\, z \in Z.$$

Therefore, if $\hat{I}$ is an isolating neighborhood for F', then it is also an isolating neighborhood for $H(\cdot, s)$ for all $s \in [0,1]$. By the continuation property of the Conley index, $\chi_*(\hat{I}, H(\cdot,1)) \cong \chi_*(\hat{I}, g')$.

Define $q : W \to X_m$ by $q(w) = P_m H(w + v_0, 1)$. By [19, Theorem 1.12], $\chi_*(I,q) \cong \chi_*(\hat{I}, H(\cdot,1))$. Finally, the result follows by [12, Theorem 5.4]. ■

The following two corollaries, which are just special cases of the previous theorem, are used in this paper. The first allows us to lift the information from $\mathbb{R}^m$ to $\mathbb{R}^{m+1}$. As was indicated in the introduction, computing the homology index is by far the most expensive calculation, and the expense grows rapidly with dimension. Therefore, we adopt the strategy of computing the index with a very coarse low-dimensional approximation, i.e., $\chi_*(I,g)$ in the following corollary. To improve the accuracy, we need to increase the dimension of the approximation. This corollary guarantees that the index does not change.

Corollary 2.4. *Let $F' : W \times [a^-, a^+] \rightrightarrows X_m \times [a^-, a^+]$ and $F : W \rightrightarrows X_m$ be multivalued maps satisfying (2.2). Let $F'_m(z) := Q_m F'(z)$ and assume that $F'_m(Z) \subset (a^-, a^+)$.*

(a) *If I is an isolating neighborhood for F, then $\hat{I} := I \times [a^-, a^+]$ is an isolating neighborhood for F'.*

(b) *If $N = (N_1, N_0)$ is an index pair for I under F, then $\hat{N} := (N_1 \times [a^-, a^+], N_0 \times [a^-, a^+])$ is an index pair for $\hat{I}$ under F'.*

(c) *Let g and g' be continuous selectors for F and F', respectively; then*

$$\chi_*(\hat{I}, g') \cong \chi_*(I, g).$$

Once the dimension of the approximation has been increased sufficiently to obtain the desired accuracy, the following result implies that the index computed with the finite-dimensional approximation is actually valid for the full infinite-dimensional system. We now return to the notation of the introduction. In particular, as in (1.10) $Z = W \times \prod_{k=m}^{\infty} \widetilde{a}_k$, where $\widetilde{a}_k := [-\frac{A_s}{|k|^s}, \frac{A_s}{|k|^s}]$ or $\widetilde{a}_k := [-\frac{A_s}{s^{|k|}}, \frac{A_s}{s^{|k|}}]$. Furthermore, s is chosen large enough that Z is compact. Also, as in (1.5), $a'_k = f_k(a)$ for all $a \in Z$.

Corollary 2.5. *Let $f : Z \to X_m \times V$ be a continuous map. Let $F : W \rightrightarrows \mathbb{R}^m$ be a multivalued map satisfying*

$$P_m f(a) \subset F(P_m a) \qquad \forall\, a \in Z.$$

Finally, assume that for all $k \geq m$ and all $a \in Z$,

$$f_k(a) \in (a_k^-, a_k^+). \tag{2.4}$$

(a) *If I is an isolating neighborhood for F, then $\hat{I} := I \times V$ is an isolating neighborhood for f.*
(b) *If $N = (N_1, N_0)$ is an index pair for I under F, then $\hat{N} := (N_1 \times V, N_0 \times V)$ is an index pair for $\hat{I}$ under f.*
(c) *If g is continuous selector for F, then*

$$\chi_*(\hat{I}, f) \cong \chi_*(I, g).$$

2.2. Dynamics via the Conley index. We now turn to the question of how the Conley index can be used to draw conclusions about the dynamics of a continuous map f. For a particularly clear and complete explanation the reader is referred to [23], which is the basis for the following discussion. Let $N = (N_1, N_0)$ be an index pair for an isolating neighborhood I of f. Let $(N_1/N_0, [N_0])$ be the pointed topological space obtained by collapsing N_0 to a single point. Since $N = (N_1, N_0)$ is an index pair, the induced quotient map

$$f_N : (N_1/N_0, [N_0]) \to (N_1/N_0, [N_0])$$

is continuous, and hence defines a dynamical system on a compact set. In particular, we have the induced map on homology,

$$f_{N*} : H_*(N_1/N_0, [N_0]) \to H_*(N_1/N_0, [N_0]).$$

We will write

$$f_{N,n} : H_n(N_1/N_0, [N_0]) \to H_n(N_1/N_0, [N_0])$$

when we need to indicate the homology map on the nth level.

Since in our applications the N_i's are polygons, f_{N*} is equivalent to the Conley index map $\chi_*(I, f)$. For the same reason, throughout this section we will assume that $H_*(N_1, N_0)$ are *free* Abelian groups; in other words, the relative homology groups of the index pairs do not have a torsion subgroup.

The simplest dynamical result is the Ważewski principle: If f_{N*} is not nilpotent, then $\mathrm{Inv}(I, f) \neq \emptyset$. Under appropriate conditions it is also fairly simple to check for the existence of a fixed point. Let

$$\Lambda(I, f) := \sum_{n=0}^{\infty} (-1)^n \mathrm{tr}\, f_{N,n}, \tag{2.5}$$

and if $\Lambda(I, f) \neq 0$, then f has a fixed point in I [23, Corollary 1.2]. Similarly, if $\Lambda(I, f^s) \neq 0$ for some positive integer s, then I contains a periodic point, though the minimal period may be less than s.

We are, of course, interested in more complicated dynamics. The simplest criterion is due to [2], where it is shown that if the spectral radius of $f_{N,1} : H_1(N_1/N_0, [N_0]) \to H_1(N_1/N_0, [N_0])$ is greater than 1, then the entropy of $\mathrm{Inv}(I, f)$ is positive. To obtain a more detailed description of the dynamics, we need to impose further restrictions on the index pair (N_1, N_0). Following the ideas of [23, 22, 21, 20], assume that the sets N_i are cubical and that

$$\mathrm{cl}\,(N_1 \setminus N_0) = \bigcup_{j=1}^{J} B_j,$$

where the B_j's are compact disjoint sets. Let $E_j := \mathrm{cl}\,(N_1 \setminus B_j)$. Then there is a continuous induced map $f_{B_j} : (N_1/E_j, [E_j]) \to (N_1/E_j, [E_j])$. Observe that if $\pi_j : (N_1/N_0, [N_0]) \to (N_1/E_j, [E_j])$ denotes the projection map, then $f_{B_j} = \pi_j \circ f_N$. Thus we can recover $f_{B_j *}$ from f_{N*}.

This information can be used to describe the invariant set in terms of symbolic dynamics as follows. Let $\Sigma_J = \{1, \dots, J\}^{\mathbb{N}}$ with the product topology. Define $\rho : \mathrm{Inv}(I, f) \to \Sigma_J$ by

$$(\rho(x))_n = j \quad \text{if} \quad f^n(x) \in B_j.$$

Let $\gamma^{t+1} = (\gamma_0, \dots, \gamma_t) \in \{1, \dots, J\}^{t+1}$. Define

$$f_{\gamma^{t+1}} = f_{B_{\gamma_t}*} \circ \cdots \circ f_{B_{\gamma_1}*}.$$

Let

$$\Gamma := \mathrm{cl}\left(\bigcup_{t \in \mathbb{N}} \left\{ \gamma^{t+1} \in \{1, \dots, J\}^{t+1} \mid f^k_{\gamma^{t+1}} \neq 0 \,\forall\, k \in \mathbb{N} \right\} \right) \subset \Sigma_J.$$

Then,

$$\Gamma \subset \rho(\mathrm{Inv}(I, f)). \tag{2.6}$$

3. Reduction of Φ.

In this section we are going to describe in detail how to reduce the infinite-dimensional map Φ to a combinatorial multivalued map $\mathcal{F} : \mathcal{G} \rightrightarrows \mathcal{G}$ on some finite set $\mathcal{G}$, which we can deal with on the computer. This reduction process involves three main steps, as outlined in the introduction. First, we recast the map into a (single-valued) countable system f of maps using a Galerkin projection. Second, we define a finite-dimensional multivalued map F in such a way that the dynamics of f is captured by F in a given subset of the phase space. Finally, we discretize the phase space of F using a finite cubical grid $\mathcal{G}$ and define a map $\mathcal{F}$ on $\mathcal{G}$, which captures the dynamics of F.

3.1. The Galerkin projection.

As laid out in the introduction, using Fourier modes $\varphi_k(x) = e^{ikx}$, $k \in \mathbb{Z}$, as the basis for L^2 we get the countable system of maps

$$a_k \mapsto f_k((a_i)_{i \geq 0}) := \mu b_k \left[a_k - \sum_{j+l+n=k} c_j a_l a_n \right], \quad k = 0, 1, 2, \dots,$$

where $a_k = a_{-k}, b_k = b_{-k}, c_k = c_{-k} \in \mathbb{R}$, $k \in \mathbb{Z}$, are the coefficients of the Fourier expansions of a, b, and c^{-1}, respectively.

3.2. Finite-dimensional approximation. We next need to reduce the countable system f to a finite-dimensional one. The idea is to do computations using the first m coordinates of f and to incorporate the neglected terms into a fixed error, which makes the finite-dimensional map multivalued.

So we split f_k into a part which depends only on the variables $a_0, \ldots, a_{m-1}$ and a remainder term:

$$f_k((a_i)_{i\geq 0}) = f_k^{(m)}(a_0, \ldots, a_{m-1}) + f_k^{(m+)}((a_i)_{i\geq 0}), \quad k = 0, 1, 2, \ldots .$$

Now $f^{(m)} := (f_0^{(m)}, \ldots, f_{m-1}^{(m)})$ defines a map on $\mathbb{R}^m$. In order to deal with the error contributed by the $f_k^{(m+)}$ term, we consider the dynamics of f on the set

$$Z = \prod_{k=0}^{\infty} [a_k^-, a_k^+] = W \times \prod_{k=m}^{\infty} [a_k^-, a_k^+]$$

(where the $a_k^\pm$'s are chosen in such a way that Z is compact) and use interval arithmetic and the estimates in section 5 to bound $f_k^{(m+)}(Z)$, $k = 0, \ldots, m-1$, by an interval $\varepsilon_k^{(m+)}[-1, 1]$; i.e., we compute $\varepsilon_k^{(m+)} > 0$, $k = 0, \ldots, m-1$, such that

$$f_k^{(m+)}(Z) \subset \varepsilon_k^{(m+)}[-1, 1].$$

The finite-dimensional multivalued map $F^{(m)} : W \rightrightarrows \mathbb{R}^m$ is now defined as

$$F^{(m)}(a_0, \ldots, a_{m-1}) = f^{(m)}(a_0, \ldots, a_{m-1}) + \prod_{k=0}^{m-1} \varepsilon_k^{(m+)}[-1, 1].$$

Obtaining the bounds $a_k^\pm$. One way to compute initial values for the bounds $a_k^\pm$ is to run a simulation of $f^{(L)} : \mathbb{R}^L \to \mathbb{R}^L$ for some large L. That is, for some arbitrary finite set $A \subset \mathbb{R}^L$ of points and some numbers $K_0 < K_1 \in \mathbb{N}$, we compute

$$A_K := \bigcup_{k=K_0}^{K_1} \left(f^{(L)}\right)^k (A).$$

For $k = 0, \ldots, L-1$, set

$$\begin{aligned} A_k^- &:= \min\{a_k \mid (a_0, \ldots, a_{L-1}) \in A_K\} \\ A_k^+ &:= \max\{a_k \mid (a_0, \ldots, a_{L-1}) \in A_K\}. \end{aligned}$$

Since the values of the bounds $a_k^\pm$ will have a strong impact on how large the errors $\varepsilon_k^{(m+)}$ will be, we decide to split these into two groups: We will use *explicit bounds* for $k < M$, where $M > m$ is some constant which will be determined by inspecting the values $A_k^\pm$. For $k \geq M$ we will use a polynomial or an exponential decaying bound for $a_k^\pm$, i.e., $a_k^\pm = \pm\frac{A_s}{k^s}$, respectively,

$a_k^\pm = \pm\frac{A_s}{s^k}$, $k \geq M$. The decay rate of the $A_K^\pm$'s yields a first guess for the constants A_s and s; i.e., we choose A_s and s such that

$$[A_k^-, A_k^+] \subset \frac{A_s}{k^s}[-1,1], \quad \text{resp.,} \quad [A_k^-, A_k^+] \subset \frac{A_s}{s^k}[-1,1],$$

for $k = M, \dots, L-1$. Now choose the $a_k^\pm$ such that

$$[A_k^-, A_k^+] \subseteq [a_k^-, a_k^+] \quad \text{for } k = 0, \dots, M-1,$$

as well as

$$a_k^\pm := \pm\frac{A_s}{k^s}, \qquad \text{resp., } a_k^\pm := \pm\frac{A_s}{s^k}, \qquad \text{for } k \geq M.$$

Determining the projection dimension. We still have to determine the dimension m used in the reduction process. If we choose m too small, the errors $\varepsilon_k^{(m+)}$ will be too big and we will not be able to extract interesting dynamics on $F^{(m)}$. On the other hand, if m is chosen too big, we will be doing unnecessary computations. So roughly speaking, we will choose the smallest m such that a numerical study of $F^{(m)}$ yields interesting results. In practice this essentially involves doing some sample computations using different values of m.

3.3. Finite representation. So far, we deduced a finite-dimensional multivalued map $F^{(m)}$ from the original system Φ. The next step will be to get a finite representation of $F^{(m)}$, which we can deal with in the computer. To this end, we are partitioning the phase space W of $F^{(m)}$ into a finite number of cubical sets. A *cubical set* is a subset B of $\mathbb{R}^m$ of the form

$$B = B(c,r) = \{x \mid |x_k - c_k| \leq r_k, k = 0, \dots, m-1\},$$

where $c, r \in \mathbb{R}^m$, $r_k \geq 0$, are the *center* and *radius* of B. Note that the number of sets in a cubical partition of W grows exponentially in the dimension m of W, a fact that would render the following computational approach prohibitively expensive. For that reason we are actually not going to work with a partition of the whole of W but instead with a cubical covering of the maximal invariant set of F in W. This way the numerical effort is essentially determined by the dimension of the maximal invariant set—which typically is much smaller than m (see [4]).

Note that W is a cubical set. Let $c = (c_0, \dots, c_{m-1})$ be the center and $r = (r_0, \dots, r_{m-1})$ the radius of W; then by *bisecting* W with respect to the jth coordinate direction, one obtains two cubical sets $B^- = B(c^-, \hat{r})$ and $B^+ = B(c^+, \hat{r})$, where

$$\hat{r}_k = \begin{cases} r_k & \text{for } k \neq j, \\ r_k/2 & \text{for } k = j, \end{cases} \qquad c_k^\pm = \begin{cases} c_k & \text{for } k \neq j, \\ c_k \pm r_k/2 & \text{for } k = j. \end{cases}$$

A cubical set, which can be represented by iterating this subdivision process, will be called a *box*. Note that a binary tree represents a certain set of boxes if one assigns a coordinate direction to each level (i.e., the set of all nodes with the same distance from the root) of the tree: the root corresponds to the box W, and all nodes of a given depth (i.e., on the same level) in the tree correspond to a subset of a cubical grid on W (see [4] for a more detailed description of this approach). Denote by $\mathcal{B}_k$ the collection of all boxes represented by the nodes on depth k of a tree (where the root has depth 0). For a subset $\mathcal{B} \subset \mathcal{B}_k$ let $|\mathcal{B}|$ denote the union of all boxes in $\mathcal{B}$. Let $o(\mathcal{B})$ be the set of all boxes in $\mathcal{B}_k$ which intersect $|\mathcal{B}|$, i.e., the smallest representable neighborhood of $|\mathcal{B}|$ in $\mathcal{B}_k$.

The multivalued map $F^{(m)}$ defines in a natural way a multivalued map $\mathcal{F}^{(m)}$ on $\mathcal{B}_k$: For $B \in \mathcal{B}_k$ let $\mathcal{F}^{(m)}(B)$ be the set of all boxes in $\mathcal{B}_k$ which intersect the set $F^{(m)}(B)$. However, in order to allow for errors introduced when computing and representing $F^{(m)}(B)$, we will actually deal with an *enclosure* of $F^{(m)}$, i.e., a map $\mathcal{F}^{(m)} : \mathcal{B}_k \rightrightarrows \mathcal{B}_k$ such that

$$F^{(m)}(B) \subset \operatorname{int} |\mathcal{F}^{(m)}(B)|$$

for $B \in \mathcal{B}_k$. It is the map $\mathcal{F}^{(m)}$ that we are dealing with in the computer. In a concrete implementation it may be represented as, e.g., a (sparse) matrix or a graph.

Computing an enclosure of F. Let us start by considering a single-valued map $g : \mathbb{R}^m \to \mathbb{R}^m$ (for example, consider the single-valued part $f^{(m)}$ of $F^{(m)}$). An approximate method for computing all boxes B' in the collection $\mathcal{B}_k$ which intersect the image $g(B)$ of a given box $B \in \mathcal{B}_k$ is to choose a finite set T of *test points* in B and to consider all boxes B' which contain at least one of the image points $f(T)$. This is the approach originally proposed in [4, 5]. Note that, due to the hierarchical storage scheme of the boxes in a binary tree, the computational complexity of determining the box $B' \in \mathcal{B}_k$, which contains some specific image point, is only $\mathcal{O}(k)$. In general this approach will not yield an enclosure of g. However, in [10] it has been shown how to extend this approach to compute an enclosure by constructing an appropriate mesh of test points in each box.

The approach used in this paper for the multivalued map is still different and based on the following observations:

1. For small enough boxes (i.e., large k), the image of a box $B \in \mathcal{B}_k$ under a map $g : \mathbb{R}^m \to \mathbb{R}^m$ is approximately given by its image under the linear part of g.
2. For a given cubical set $C \subset \mathbb{R}^m$, the set of all boxes $B' \in \mathcal{B}_k$ which intersect C can efficiently be determined by a single depth first search in the tree.

So the idea is to use the linear part of $f^{(m)}$ (the single-valued part of $F^{(m)}$) to compute an approximate image of a box B, to enclose this image by a cubical set, and then to enlarge this cubical set by the errors made by neglecting the nonlinear terms of $f^{(m)}$, as well as the multivalued part of $F^{(m)}$. In doing these computations we use interval arithmetic as implemented in the BIAS, PROFIL, and `b4m` libraries (see [13, 14, 25]) in order to control round-off errors.

Let us be more precise. Consider the box $B = B(c, r) \in \mathcal{B}_k$. For $h \in \mathbb{R}^m$ we can decompose $f^{(m)}$ as

$$f^{(m)}(c + h) = f^{(m)}(c) + Df^{(m)}(c)h + f^{(m),nl}(c, h),$$

where $f^{(m),nl}(c, 0) = 0$. Compute $\varepsilon^{(m),nl}(c) \in \mathbb{R}^m$ such that

$$\max_{h \in B(0,r)} |f^{(m),nl}(c, h)| \leq \varepsilon^{(m),nl}(c).$$

Then $F^{(m)}(B)$ will be contained in the cubical set $B(f^{(m)}(c), R)$, where

$$R = |Df^{(m)}(c)|r + \varepsilon^{(m),nl}(c) + \varepsilon^{(m+)},$$

and for a matrix $A = (a_{ij}) \in \mathbb{R}^{d,e}$ we write $|A| := (|a_{ij}|) \in \mathbb{R}^{d,e}$. One should emphasize that the computation of $\varepsilon^{(m),nl}$ may eventually be expensive—it is not in our case. Now the enclosure $\mathcal{F} : \mathcal{B}_k \rightrightarrows \mathcal{B}_k$ of $F^{(m)}$ is defined in the following way: Let $\mathcal{F}(B)$ be the set of boxes which is intersected by the cubical set $B(f^{(m)}(c), R)$.

The following algorithm (when called as $\texttt{cap}(\emptyset, W, C, k)$) computes the set $\mathcal{I}$ of all boxes in $\mathcal{B}_k$ which have a nonempty intersection with the cubical set C.

Algorithm 1.

```
$\mathcal{I}$ = cap($\mathcal{I}, B, C, k$)
     if $B \cap C \neq \emptyset$
          if depth($B$) $= k$
               $\mathcal{I} := \mathcal{I} \cup \{B\}$
          else
               $\mathcal{I} := \mathcal{I} \cup$ cap($\mathcal{I}, B^+, C, k$) $\cup$ cap($\mathcal{I}, B^-, C, k$)
     return $\mathcal{I}$
```

Here the function $\texttt{depth}(B)$ returns k if $B \in \mathcal{B}_k$, and B^+ and B^- are the two boxes which result from bisecting B with respect to some coordinate direction—as defined by the tree used to store the box collections.

4. Computing isolating neighborhoods.

After reducing the map Φ to a combinatorial map $\mathcal{F}$ on a grid, we are now interested in computing isolating neighborhoods for $\mathcal{F}$ which isolate certain sets of interest. In particular we will be interested in isolating periodic points, connecting orbits and—combining these two—invariant sets with complicated dynamics. These neighborhoods will translate directly into isolating neighborhoods for the multivalued map F (see section 1) and—together with certain conditions on f—into isolating neighborhoods for f and thus for Φ (see section 2).

Szymczak [22] describes an algorithm for finding isolating neighborhoods of $\mathcal{F}$ in a given subset $\mathcal{B}$ of $\mathcal{B}_k$. The basic idea is to "cut" the proper pieces out of $\mathcal{B}$ until the resulting collection satisfies the criterion (1.12). This method has the drawback that one has to choose a suitable set $\mathcal{B}$ a priori and may eventually end up with the empty set as a trivial isolating neighborhood. The approach we will describe now proceeds in some sense in the opposite direction: one starts with a *guess* for an isolating neighborhood of some interesting invariant set and "fattens" this set by adding neighborhoods until condition (1.12) is satisfied.

4.1. Guessing isolating neighborhoods.

Guesses for isolating neighborhoods of specific invariant sets of $\mathcal{F}$ can easily be obtained as follows:

- k-periodic points of $\mathcal{F}$ are identified by nonzero diagonal entries of $M_{\mathcal{F}}^k$, where the *transition matrix* $M_{\mathcal{F}} = (m_{ij})$ is given by
$$m_{ij} = \begin{cases} 1 & \text{if } B_j \in \mathcal{F}(B_i), \\ 0 & \text{else,} \end{cases}$$
and $\mathcal{B}_k = \{B_1, \ldots, B_p\}$;
- more generally, recurrent sets of $\mathcal{F}$ are given by strongly connected components of a graph representing $\mathcal{F}$;
- in this graph, connecting orbits of $\mathcal{F}$ can be identified by shortest path algorithms (e.g., the Dijkstra algorithm; see [6]);

- Szymzcak [22] describes how to compute the maximal invariant set of $\mathcal{F}$.

4.2. Turning the guess into a true isolating neighborhood. Once a guess $\tilde{\mathcal{I}}$ for an isolating neighborhood of $\mathcal{F}$ has been computed, we construct by the following procedure a (true) isolating neighborhood $\mathcal{I}$ containing $\tilde{\mathcal{I}}$.

Algorithm 2.

```
$\mathcal{I}$ = make_isolated($\tilde{\mathcal{I}}$)
    $\mathcal{I}$ := $\mathrm{Inv}(\tilde{\mathcal{I}}, \mathcal{F})$
    while $o(\mathcal{I}) \not\subset \tilde{\mathcal{I}}$
        $\tilde{\mathcal{I}}$ := $\tilde{\mathcal{I}} \cup o(\mathcal{I})$
        $\mathcal{I}$ := $\mathrm{Inv}(\tilde{\mathcal{I}}, \mathcal{F})$
    if $\mathcal{I} \subset \mathrm{int}\,|o(\mathcal{I})|$ return $\mathcal{I}$
    else return $\emptyset$
```

By construction this algorithm returns a combinatorial isolating neighborhood $\mathcal{I}$ for $\mathcal{F}$. Similarly to the procedure proposed in [22], one may end up with the empty set, in which case the set $|\mathcal{I}|$ touched the boundary of W.

4.3. Tightening the isolating neighborhood. So far we computed an isolating neighborhood $\mathcal{I} \subset \mathcal{B}_k$ for $\mathcal{F}$. We are now going to address the question of how to improve this neighborhood in the sense that one gets a tighter covering of the underlying invariant set. This process of tightening involves three steps, which may be applied repeatedly until the desired accuracy has been reached or the machine resources are exhausted.

4.3.1. The subdivision algorithm. In [4] Dellnitz and Hohmann describe a subdivision procedure for the computation of *relative global attractors* of maps. The basic idea of the therein advocated multilevel approach is to iteratively refine a given collection of boxes and then to select a certain subset of the refined collection which contains the dynamics of interest.

Algorithm 3 (see [4]). *Given the initial collection $\mathcal{B}_0$, one inductively obtains $\mathcal{B}_k$ from $\mathcal{B}_{k-1}$ for $k = 1, 2, \dots$ in two steps.*

1. Subdivision: *Construct a new collection $\hat{\mathcal{B}}_k$ by bisecting each box in $\mathcal{B}_{k-1}$ with respect to some coordinate direction.*
2. Selection: *Compute the relevant subset $\mathcal{B}_k$ of $\hat{\mathcal{B}}_k$.*

The second step obviously defines which sets will be contained in $\mathcal{B}_k$. In our case, since we want to compute isolating neighborhoods, we are interested in covering the maximal invariant set. So we set

$$\mathcal{B}_k = \mathrm{Inv}(\hat{\mathcal{B}}_k, \mathcal{F}) \tag{4.1}$$

(cf. [22]), where we start with $\mathcal{B}_0 = \mathcal{I}$. The following statement formalizes that we do not lose the isolation property for the tightened neighborhood. Its proof is essentially the same as that for Theorem 2.2 in [22].

Proposition 4.1. *Let $\mathcal{B}_0 = \mathcal{I}$ be a collection of boxes which is an isolating neighborhood for $\mathcal{F} : \mathcal{B}_0 \rightrightarrows \mathcal{B}_0$. Then the collections $\mathcal{B}_k$, $k = 1, 2, \dots,$ computed by Algorithm 3 with selection criterion (4.1), are also isolating neighborhoods for $\mathcal{F} : \mathcal{B}_k \rightrightarrows \mathcal{B}_k$.*

In practice it will not be advisable to perform more than a few steps of the subdivision procedure at once. Recall that the combinatorial map $\mathcal{F}$ has been defined on the basis of the multivalued map $F = F^{(m)}$, which incorporated the errors $\varepsilon_k^{(m+)}$ contributed to the dynamics of $f^{(m)}$ by the higher order modes. As soon as these errors and the size of the boxes in the current collection $\mathcal{B}_k$ are on the same order of magnitude, a further refinement using Algorithm 3 no longer makes any sense. Instead, one will first have to make $\varepsilon_k^{(m+)}$ smaller by the following two methods.

4.3.2. Updating the bounds $a_k^\pm$. Recall that we start all computations on the infinite-dimensional cube

$$Z = W \times V = \prod_{k=0}^{\infty} [a_k^-, a_k^+],$$

where we obtained initial guesses for $a_k^\pm$ by running a simulation of $f^{(m)}$ for a large m. The size of $a_k^\pm$ will have a crucial influence on the size of the errors $\varepsilon_k^{(m)+}$ introduced in the multivalued map $F^{(m)}$ by estimating the contribution from the neglected modes. Since these errors determine the precision of the computations (and whether we will be able to perform an interesting computation in the first place) we are interested in making $|a_k^\pm|$ as small as possible. Remember that we split the variables a_k into three groups as follows:

1. $0 \leq k < m$: These are the actual variables with which we are computing. We are getting tight bounds on these by encapsulating our covering $\mathcal{B}_\ell$ of the maximal invariant set of $F^{(m)} : W \rightrightarrows \mathbb{R}^m$ into a cube; i.e., we choose $a_k^\pm$, $0 \leq k < m$, such that

$$|\mathcal{B}_\ell| \subset \prod_{k=0}^{m-1} [a_k^-, a_k^+].$$

2. $m \leq k < M$: For these we store bounds on the a_k explicitly. As laid out in section 2, in order to lift the index information to the full system it is sufficient to require that the map f satisfy

$$f_k(Z) \subset (a_k^-, a_k^+), \quad k \geq m.$$

   Note that, via the estimates in section 5, we are able to bound $f_k(Z)$ in terms of an interval. What is more, since we are dealing with a polynomial nonlinearity, whenever we decrease $|a_\ell^\pm|$ for some ℓ, the bound on $f_k(Z)$ will also decrease. This is the basis for the following update scheme for the $a_k^\pm$ with $m \leq k < M$: For $k = m, \dots, M-1$ we compute the new $[a_k^-, a_k^+]$ as the interval bounding $f_k(Z)$ using the estimates in section 5, in particular, Corollary 5.5, respectively, 5.11.
3. $k \geq M$: These variables are bounded by a decay law of exponential or polynomial type, i.e., $a_k^\pm = \pm\frac{A_s}{s^k}$, respectively, $a_k^\pm = \pm\frac{A_s}{k^s}$. In this case one gets tighter bounds by updating the constants s and A_s. This is detailed in section 5 (Lemmas 5.7 and 5.13).

4.3.3. Increasing the projection dimension m. The third method for obtaining tighter bounds on the computed invariant set is to increase the number of Galerkin modes used in the computations. We write $\mathcal{B}_k = \mathcal{B}_k^{(m)}$ for a box collection in $\mathbb{R}^m$. Define

$$\mathcal{B}_k^{(m+1)} = \left\{ B \times [a_m^-, a_m^+] : B \in \mathcal{B}_k^{(m)} \right\}$$

and let $\mathcal{F}^{(m+1)} : \mathcal{B}_k^{(m+1)} \rightrightarrows \mathcal{B}_k^{(m+1)}$ be the combinatorial map defined by the multivalued map $F^{(m+1)} : W^{(m+1)} \rightrightarrows \mathbb{R}^{m+1}$, $W^{(m+1)} = W^{(m)} \times [a_m^-, a_m^+]$. Now let $\mathcal{I}^{(m)} \subset \mathcal{B}_k^{(m)}$ be an isolating neighborhood for $\mathcal{F}^{(m)} : \mathcal{B}_k^{(m)} \rightrightarrows \mathcal{B}_k^{(m)}$ and set

$$\mathcal{I}^{(m+1)} = \{B \times [a_m^-, a_m^+] : B \in \mathcal{I}^{(m)}\}.$$

Using Corollary 2.4 we get that $\mathcal{I}^{(m+1)}$ is an isolating neighborhood for $\mathcal{F}^{(m+1)}$, and in particular, the index information of a corresponding index pair carries over from dimension m to $m+1$.

5. Error bounds (polynomial nonlinearity).

As discussed in section 1, we need to study maps of the form

$$(a_k)_k \mapsto b_k \sum_{n_0,\dots,n_{p-1}\in\mathbb{Z}} c_{n_0}^p a_{n_1} \dots a_{n_{p-1}} a_{k-(n_0+\dots+n_{p-1})}$$

corresponding to a monomial $c_p a^p$ of the growth function, where $a_k = a_{-k}$, b_k is the kth eigenvalue for the underlying linear operator, and c_k^p and a_k are the kth coefficients of the expansions of c_p and a, respectively. The maps for a polynomial growth function are the sums of the maps for the monomials. It is important to note, however, that when written in terms of an appropriate basis, a wide variety of systems with polynomial nonlinearities produce maps of this form.

We next restrict the domain to a set $Z = \prod_k [a_k^-, a_k^+] = W \times \prod_{k\geq M}[a_k^-, a_k^+]$, where for $k \geq M$, $0 \in [a_k^-, a_k^+]$, and $a_k^+ - a_k^-$ satisfies some decay rule. The justification of this restriction is given by the following lemma.

Lemma 5.1. *Any invariant set of Φ is contained in a set Z of the form given above, where the decay in the higher modes reflects the decay of the eigenvalues b_k of the linear operator.*

Proof. Let $a \in X$ with corresponding Fourier expansion $\sum_k a_k \varphi_k$.

The projection of the image of a onto the kth mode is

$$\begin{aligned}
\langle \Phi(a), \varphi_k \rangle &= \int_{-\pi}^{\pi} \Phi(a)(y)\varphi_k(y)dy \\
&= \int_{-\pi}^{\pi} \frac{1}{2\pi} \int_{-\pi}^{\pi} b(x,y) g(a)(x) \varphi_k(y) dx dy \\
&= \frac{1}{2\pi} \sum_n b_n \int_{-\pi}^{\pi} \varphi_n(x) g[a](x) \left(\int_{-\pi}^{\pi} \varphi_n(-y)\varphi_k(y)dy \right) dx \\
&= b_k \int_{-\pi}^{\pi} \varphi_k(x) g[a](x) dx \\
&= b_k \langle g[a], \varphi_k \rangle \\
&\leq C_{g,a} b_k
\end{aligned}$$

for some constant $C_{g,a}$ that does not depend on k.

In particular, any set which is invariant (forward and backward in time) must be contained in a set of the form of $Z = \prod_k [a_k^-, a_k^+]$, where for k sufficiently large ($k \geq M$), $a_k^+ - a_k^-$ is shrinking according to the contraction given by the eigenvalue b_k. ■

In practice, the domain intervals are determined by preliminary simulations, with a decay rule reflecting the decay found in the eigenvalues for the linear operator. We now use interval arithmetic to determine bounds for the error term $f^{(m+)}(a)$ resulting from considering only the first m modes as variables. To emphasize our use of intervals we will write $\tilde{a}_k := [a_k^-, a_k^+]$.

5.1. Exponential decay. In this section, we assume that the eigenvalues b_k of the linear operator decay exponentially. That is, there exist constants $b > 1$ and $B > 0$ such that $|b_k| \leq \frac{B}{b^k}$. In this case, we begin by setting $\tilde{a}_k = \left[-\frac{A_s}{s^k}, \frac{A_s}{s^k}\right]$ for all $k \geq M$, where the constants $A_s > 0$ and $s > 1$ are determined by preliminary simulations. In order to reduce the number of cases we need to consider, we will extend the asymptotic bounds to all k. In other words, $\tilde{a}_k \subseteq \frac{A}{s^{|k|}}[-1,1]$ for all $k \in \mathbf{Z}$, where $A = \max\{A_s, \max_{0 \leq k < M, a_k \in \tilde{a}_k} s^k |a_k|\}$. For future computations, fix β such that $\frac{b}{s} < \beta < b$.

Consider the monomial a^p of order p with coefficient 1. We begin by bounding the sum in the corresponding maps.

Lemma 5.2. *For all $k \in \mathbb{Z}$,*

$$\sum_{n_1,\ldots,n_{p-1} \in \mathbb{Z}} \tilde{a}_{n_1} \ldots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \subseteq \frac{\alpha^{p-1} A^p}{s^{|k|}} \left(\frac{b}{\beta}\right)^{|k|} [-1,1],$$

where $\alpha = \left(\frac{2}{\ln s} + \frac{b}{\beta \ln (b/\beta)}\right)$.

Proof. This formula holds for $p = 1$ since

$$\tilde{a}_k \subseteq \frac{A}{s^{|k|}}[-1,1] \subseteq \frac{A}{s^{|k|}} \left(\frac{b}{\beta}\right)^{|k|} [-1,1].$$

Assume the formula holds for $p-1$. Then, for $k \geq 0$,

$$\begin{aligned}
&\sum_{n_1,\ldots,n_{p-1}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\quad = \sum_{n_1} \tilde{a}_{n_1} \sum_{n_2,\ldots,n_{p-1}} \tilde{a}_{n_2} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{(k-n_1)-(n_2+\cdots+n_{p-1})} \\
&\quad \subseteq \sum_{n_1} \tilde{a}_{n_1} \frac{\alpha^{p-2} A^{p-1}}{s^{|k-n_1|}} \left(\frac{b}{\beta}\right)^{|k-n_1|} [-1,1] \\
&\quad \subseteq \left(\sum_{n_1 \leq k} \tilde{a}_{n_1} \frac{\alpha^{p-2} A^{p-1}}{s^{k-n_1}} \left(\frac{b}{\beta}\right)^{k-n_1} + \sum_{n_1 > k} \tilde{a}_{n_1} \frac{\alpha^{p-2} A^{p-1}}{s^{n_1-k}} \left(\frac{b}{\beta}\right)^{n_1-k} \right) [-1,1] \\
&\quad \subseteq \left(\sum_{n_1 > 0} \frac{A\alpha^{p-2} A^{p-1}}{s^{k+2n_1}} \left(\frac{b}{\beta}\right)^{k+n_1} + \sum_{n_1=0}^{k} \frac{A\alpha^{p-2} A^{p-1}}{s^k} \left(\frac{b}{\beta}\right)^{k-n_1} \right. \\
&\qquad \left. + \sum_{n_1 > k} \frac{A\alpha^{p-2} A^{p-1}}{s^{2n_1-k}} \left(\frac{b}{\beta}\right)^{n_1-k} \right) [-1,1]
\end{aligned}$$

$$
\begin{aligned}
&\subseteq \left(\frac{\alpha^{p-2}A^p}{s^k} \left(\frac{b}{\beta}\right)^k \left(\frac{b}{s\beta}\right)^0 \sum_{n_1>0} \frac{1}{s^{n_1}} + \frac{\alpha^{p-2}A^p}{s^k} \left(\frac{b}{\beta}\right)^k \sum_{n_1=0}^{k} \left(\frac{\beta}{b}\right)^{n_1} \right.\\
&\qquad \left. + \alpha^{p-2}A^p s^k \left(\frac{\beta}{b}\right)^k \left(\frac{b}{s\beta}\right)^k \sum_{n_1>k} \frac{1}{s^{n_1}} \right)[-1,1]\\
&\subseteq \left(\frac{\alpha^{p-2}A^p}{s^k} \left(\frac{b}{\beta}\right)^k \int_0^\infty s^{-x} dx + \frac{\alpha^{p-2}A^p}{s^k} \left(\frac{b}{\beta}\right)^k \int_{-1}^\infty \left(\frac{\beta}{b}\right)^x dx \right.\\
&\qquad \left. + \alpha^{p-2}A^p \int_k^\infty s^{-x} dx \right)[-1,1]\\
&\subseteq \frac{\alpha^{p-2}A^p}{s^k} \left(\frac{b}{\beta}\right)^k \left(\frac{2}{\ln s} + \frac{b}{\beta \ln(b/\beta)}\right)[-1,1]\\
&= \frac{\alpha^{p-1}A^p}{s^k} \left(\frac{b}{\beta}\right)^k [-1,1].
\end{aligned}
$$

The case $k < 0$ may be reduced to the previous case via a change of indices. ■

If the expansion of the coefficient function exhibits similar decay to that of $\tilde{a}_k$, then we may extend this argument to the maps corresponding to ca^p.

Corollary 5.3. *If there exists a constant C such that $c_n \in \tilde{c}_n := \frac{C}{s^{|n|}}[-1,1]$ for all n, then*

$$
\sum_{n_0,n_1,\dots,n_{p-1}\in\mathbf{Z}} \tilde{c}_{n_0}\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_0+\cdots+n_{p-1})} \subseteq \frac{\alpha^p A^p C}{s^k}\left(\frac{b}{\beta}\right)^k [-1,1].
$$

We now take advantage of the explicit bounds $\tilde{a}_n$, $0 \le n < \overline{M}$, instead of using only the extended asymptotic bounds $\frac{A_s}{s^n}[-1,1]$. By increasing $\overline{M}$, the error computations become more costly while giving tighter bounds on the error terms.

Lemma 5.4. *For $0 \le k < \overline{M}$,*

$$
\begin{aligned}
&\sum_{n_1,\dots,n_{p-1}\in\mathbb{Z}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})}\\
&\qquad\subseteq \sum_{|n_1|,\dots,|n_{p-1}|\le\overline{M}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})}\\
&\qquad\qquad + \frac{(p-1)\alpha^{p-2}A^{p-1}A_s}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k} + \left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right][-1,1]
\end{aligned}
$$

for any $\overline{M} > 0$.

Proof.

$$\begin{aligned}
\sum_{n_1,\ldots,n_{p-1}\in\mathbb{Z}} & \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\subseteq \sum_{|n_1|,\ldots,|n_{p-1}|\leq M^*} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\qquad + \sum_{i=1}^{p-1}\max_{a_n\in\tilde{a}_n}\left\{\sum_{n_j\in\mathbb{Z},|n_i|>\overline{M}} |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\ldots+n_{p-1})}|\right\}[-1,1].
\end{aligned}$$

By symmetry,

$$\sum_{n_j\in\mathbb{Z},|n_i|>\overline{M}} |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}| = \sum_{n_j\in\mathbb{Z},|n_1|>\overline{M}} |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}|$$

for $i = 1,\ldots,p-1$.

Using the previous asymptotic estimates,

$$\begin{aligned}
\sum_{n_j\in\mathbb{Z},|n_1|>\overline{M}} & |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}| \\
&\leq \sum_{n_1>\overline{M}} \frac{A_s}{s^{n_1}}\frac{\alpha^{p-2}A^{p-1}}{s^{n_1-k}}\left(\frac{b}{\beta}\right)^{n_1-k} \\
&\qquad + \sum_{n_1<-\overline{M}} \frac{A_s}{s^{-n_1}}\frac{\alpha^{p-2}A^{p-1}}{s^{k-n_1}}\left(\frac{b}{\beta}\right)^{k-n_1} \\
&= \alpha^{p-2}A^{p-1}A_s s^k\left(\frac{\beta}{b}\right)^k\left(\frac{b}{s\beta}\right)^{\overline{M}}\sum_{n_1>\overline{M}}\frac{1}{s^{n_1}} \\
&\qquad + \sum_{n_1>\overline{M}} \frac{A_s}{s^{n_1}}\frac{\alpha^{p-2}A^{p-1}}{s^{k+n_1}}\left(\frac{b}{\beta}\right)^{k+n_1} \\
&\leq \alpha^{p-2}A^{p-1}A_s s^k\left(\frac{\beta}{b}\right)^k\left(\frac{b}{s\beta}\right)^{\overline{M}}\int_{\overline{M}}^{\infty} s^{-x}dx \\
&\qquad + \frac{\alpha^{p-2}A^{p-1}A_s}{s^k}\left(\frac{b}{\beta}\right)^k\left(\frac{b}{s\beta}\right)^{\overline{M}}\int_{\overline{M}}^{\infty} s^{-x}dx \\
&= \frac{\alpha^{p-2}A^{p-1}A_s}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k} + \left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right].
\end{aligned}$$

Therefore,

$$
\begin{aligned}
\sum_{n_1,\dots,n_{p-1}\in\mathbb{Z}} & \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
& \subseteq \sum_{|n_1|,\dots,|n_{p-1}|\le\overline{M}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
& \quad + \frac{(p-1)\alpha^{p-2}A^{p-1}A_s}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k}+\left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right][-1,1]. \qquad \blacksquare
\end{aligned}
$$

By a similar argument, we obtain the following corollary.

Corollary 5.5. *For* $0\le k<\overline{M}$*,*

$$
\begin{aligned}
\sum_{n_0,\dots,n_{p-1}\in\mathbb{Z}} & \tilde{c}^p_{n_0}\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
& \subseteq \sum_{|n_0|,\dots,|n_{p-1}|\le\overline{M}} \tilde{c}^p_{n_0}\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
& \quad + \frac{\alpha^{p-1}A^pC}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k}+\left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right][-1,1] \\
& \quad + \frac{(p-1)\alpha^{p-1}A^{p-1}CA_s}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k}+\left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right][-1,1]
\end{aligned}
$$

for any $\overline{M}>0$.

Notice that the sum in these last estimates is finite. Therefore, we may decide on a case-by-case basis which of these terms contains only variables (a_k with $0\le k<m$). These terms are included in the finite-dimensional map, $f^{(m)}$, and should not be considered when bounding the error term, $f^{(m+)}$. In the remaining terms, we may use the explicit interval bounds instead of the extended asymptotic bounds. This, in principle, should give us a tighter bound on the error.

Corollary 5.6. *For* $0\le k<m$ *and* $\overline{M}>0$*, the error in the* k*th coordinate map,* $[f^{(m+)}(Z)]_k$*, corresponding to a nonlinear term of the form* c_pa^p *is bounded by*

$$
\begin{aligned}
[f^{(m+)}(Z)]_k \subseteq & \sum_{|n_0|,\dots,|n_{p-1}|\le\overline{M}} \overline{c^p_{n_0}a_{n_1}\cdots a_{n_{p-1}}a_{k-(n_0+\cdots+n_{p-1})}} \\
& + (A+(p-1)A_s)\frac{(\alpha A)^{p-1}C}{s^{\overline{M}}\ln s}\left[\left(\frac{b}{s\beta}\right)^{\overline{M}-k}+\left(\frac{b}{s\beta}\right)^{\overline{M}+k}\right][-1,1],
\end{aligned}
$$

where $\overline{c^p_{n_0}a_{n_1}\cdots a_{n_{p-1}}a_{k-(n_0+\cdots+n_{p-1})}}$ *is* 0 *if all of the indices have absolute value less than* m *and is the interval product* $\tilde{c}^p_{n_0}\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_0+\cdots+n_{p-1})}$ *otherwise.*

Besides computing the errors for the multivalued map, we also wish to update both the explicit interval bounds ($0\le k<M$) and the asymptotic bounds for the tail of the sequence $k\ge M$. For the first set of updates, ($0\le k<M$), we may use the estimates given in Corollary

5.5. When updating the bounds for $k \geq M$, we use the computations given in Corollary 5.3 as follows.

Lemma 5.7. *Suppose the nonlinearity is $g(a) = \sum_{p=0}^{d} c_p a^p$ with the expansions of the coefficient functions satisfying the decay rules $c_n^p \in \tilde{c}_n^p := \frac{C_p}{s^{|n|}}[-1,1]$ for some constants C_p. Then for $k \geq M$, we may set new bounds $\tilde{a}_k$ to be $\frac{A_{\beta s}}{(\beta s)^k}[-1,1]$, where*

$$A_{\beta s} = B(C_0 + \alpha A C_1 + \cdots + \alpha^d A^d C_d)$$

Proof. For a fixed $k \geq M$, we have bounds for the image of the corresponding map using the bounds for $|b_k|$ and the bounds for the sum given by Corollary 5.3.

The projection of the image of Z under the full map onto the kth mode is

$$\begin{aligned}
[F(Z)]_k &= b_k \sum_{p=0}^{d} \sum_{n_0,\ldots,n_{p-1}} \tilde{c}_{n_0}^p \tilde{a}_{n_1} \ldots \tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
&\subseteq \frac{B}{b^k} \frac{1}{s^k} \left(\frac{b}{\beta}\right)^k (C_0 + \alpha A C_1 + \cdots + \alpha^d A^d C_d)[-1,1] \\
&= \frac{A_{\beta s}}{(\beta s)^k}[-1,1],
\end{aligned}$$

where $A_{\beta s}$ is as given in the statement of the lemma.

By setting the new bounds $\tilde{a}_k$ to be the bounds on the image $[F(Z)]_k$, we preserve isolation. ■

5.2. Polynomial decay.

We now consider the case when the eigenvalues, b_k, for the linear operator exhibit polynomial decay. In other words, there exist constants $b > 1$ and $B > 0$ such that $|b_k| \leq \frac{B}{|k|^b}$ for all $k \in \mathbf{Z} \backslash \{\mathbf{0}\}$. Again, we assume that the sequence exhibits similar decay to that of the eigenvalues. That is, for some constants $A_s > 0$ and $s > 1$ initially given by simulations, $\tilde{a}_k = \frac{A_s}{|k|^s}[-1,1]$ for all $k > M$. Set $A = \max\{A_s, \max_{a_0 \in \tilde{a}_0} |a_0|, \max_{0<k<M, a_k \in \tilde{a}_k} |k|^s |a_k|\}$. Then $\tilde{a}_k \subseteq \frac{A}{|k|^s}[-1,1]$ for all $k \in \mathbf{Z} \backslash \{\mathbf{0}\}$ and $\tilde{a}_0 \subseteq A[-1,1]$. The following estimates are similar to those given in the exponential decay case.

Lemma 5.8. *Let $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$. Then*

$$\sum_{n_1,\ldots,n_{p-1} \in \mathbb{Z}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \subseteq \begin{cases} \frac{\alpha^{p-1} A^p}{|k|^s}[-1,1], & k \neq 0, \\ \alpha^{p-1} A^p [-1,1], & k = 0. \end{cases}$$

Proof. For $p = 1$, this inequality holds for all k. Now assume that

$$\sum_{n_1,\ldots,n_{p-2} \in \mathbb{Z}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-2}} \tilde{a}_{k-(n_1+\cdots+n_{p-2})} \subseteq \begin{cases} \frac{\alpha^{p-2} A^{p-1}}{|k|^s}[-1,1], & k \neq 0, \\ \alpha^{p-2} A^{p-1} [-1,1], & k = 0. \end{cases}$$

For $k = 0$,

$$\begin{aligned}
&\sum_{n_1,\dots,n_{p-1}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\qquad \subseteq \left(\sum_{n_1<0} \frac{A}{(-n_1)^s}\alpha^{p-2}A^{p-1} + A\alpha^{p-2}A^{p-1}\right. \\
&\qquad\qquad \left. + \sum_{n_1>0} \frac{A}{n_1^s}\alpha^{p-2}A^{p-1}\right)[-1,1] \\
&\qquad \subseteq \alpha^{p-2}A^p\left[2\left(1+\int_1^\infty t^{-s}dt\right)+1\right][-1,1] \\
&\qquad = \alpha^{p-2}A^p\left[\frac{3s-2}{s-1}\right][-1,1] \\
&\qquad \subset \alpha^{p-1}A^p[-1,1].
\end{aligned}$$

The following inequality is needed for the case $k > 0$:

$$\sum_{n=1}^{k-1} \frac{1}{n^s}\frac{1}{(k-n)^s} \leq \frac{3.5\cdot 2^s}{k^s}.$$

First, note that we may assume $k > 2$ since the sum is empty when $k = 1$ and the inequality holds for the single term when $k = 2$.

For $s = 2$,

$$\begin{aligned}
\sum_{n=1}^{k-1} \frac{1}{n^2}\frac{1}{(k-n)^2} &\leq \frac{2}{(k-1)^2} + \int_1^{k-1}\frac{1}{x^2(k-x)^2}dx \\
&\leq \frac{2}{(k-1)^2} + \frac{4}{k^3}\ln|k-1| + \frac{2(k-2)}{k^2(k-1)} \\
&\leq \frac{2}{(k-1)^2} + \frac{6}{k^2} \\
&< \frac{8}{k^2} + \frac{6}{k^2} \\
&= \frac{3.5\cdot 2^2}{k^2}.
\end{aligned}$$

Here, $\frac{2}{(k-1)^2} < \frac{8}{k^2}$ since $k > 2$.

Now assume that

$$\sum_{n=1}^{k-1} \frac{1}{n^{s-1}(k-n)^{s-1}} \leq \frac{3.5\cdot 2^{s-1}}{k^{s-1}}$$

or, equivalently,

$$\sum_{n=1}^{k-1} \frac{k^{s-1}}{n^{s-1}(k-n)^{s-1}} \leq 3.5\cdot 2^{s-1}.$$

Then

$$\begin{aligned}
\sum_{n=1}^{k-1} \frac{k^s}{n^s(k-n)^s} &= \sum_{n=1}^{k-1} \left(\frac{k}{n(k-n)}\right) \frac{k^{s-1}}{n^{s-1}(k-n)^{s-1}} \\
&\le \frac{k}{k-1} \sum_{n=1}^{k-1} \frac{k^{s-1}}{n^{s-1}(k-n)^{s-1}} \\
&\le 2 \cdot 3.5 \cdot 2^{s-1} \\
&= 3.5 \cdot 2^s.
\end{aligned}$$

Therefore,

$$\sum_{n=1}^{k-1} \frac{1}{n^s} \frac{1}{(k-n)^s} \le \frac{3.5 \cdot 2^s}{k^s}$$

for all $s \ge 2$.

Now, for $k > 0$,

$$\begin{aligned}
&\sum_{n_1,\dots,n_{p-1}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\quad \subseteq \sum_{n_1} \tilde{a}_{n_1} | \frac{\alpha^{p-2} A^{p-1}}{|k-n_1|^s} [-1,1] \\
&\quad \subseteq \left(\sum_{n_1<0} \frac{A}{(-n_1)^s} \frac{\alpha^{p-2} A^{p-1}}{(k-n_1)^s} + \frac{A \alpha^{p-2} A^{p-1}}{k^s} \right. \\
&\qquad\quad \left. + \sum_{n_1=1}^{k-1} \frac{A}{n_1^s} \frac{\alpha^{p-2} A^{p-1}}{(k-n_1)^s} + \sum_{n_1>k} \frac{A}{n_1^s} \frac{\alpha^{p-2} A^{p-1}}{(n_1-k)^s} \right) [-1,1] \\
&\quad \subseteq \alpha^{p-2} A^p \left[\frac{2}{k^s} \left(1 + \int_1^\infty t^{-s} dt \right) \right. \\
&\qquad\qquad \left. + \sum_{n_1=1}^{k-1} \frac{1}{n_1^s} \frac{1}{(k-n_1)^s} \right] [-1,1] \\
&\quad \subseteq \alpha^{p-2} A^p \left[\frac{2}{k^s} \left(1 + \frac{1}{s-1} \right) + \frac{3.5 \cdot 2^s}{k^s} \right] [-1,1] \\
&\quad \subseteq \frac{\alpha^{p-2} A^p}{k^s} \left[\frac{2}{s-1} + 2 + 3.5 \cdot 2^s \right] [-1,1] \\
&\quad = \frac{\alpha^{p-1} A^p}{k^s} [-1,1].
\end{aligned}$$

The case $k < 0$ may be reduced to the previous case via a change of indices. ■

We may extend this argument to the maps corresponding to ca^p, provided that the expansion of the coefficient function also exhibits polynomial decay.

Corollary 5.9. *If there exists a constant C such that $c_n \in \tilde{c}_n := \frac{C}{|n|^s}[-1,1]$ for all n, then*

$$\sum_{n_0,n_1,\ldots,n_{p-1}\in\mathbf{Z}} \tilde{c}_{n_0}\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_0+\cdots+n_{p-1})} \subseteq \begin{cases} \frac{\alpha^p A^p C}{|k|^s}[-1,1], & k\neq 0, \\ \alpha^p A^p C[-1,1], & k=0. \end{cases}$$

We now refine the error bounds using the explicit interval bounds $\tilde{a}_n$ for $n<\overline{M}$.

Lemma 5.10. *For $0\le k<\overline{M}$,*

$$\begin{aligned}\sum_{n_1,\ldots,n_{p-1}\in\mathbb{Z}} &\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\ &\subseteq \sum_{|n_1|,\ldots,|n_{p-1}|\le\overline{M}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\ &\quad + \frac{(p-1)\alpha^{p-2}A^{p-1}A_s}{\overline{M}^{s-1}(s-1)}\left[\frac{1}{(\overline{M}-k)^s}+\frac{1}{(\overline{M}+k)^s}\right][-1,1]\end{aligned}$$

for any $\overline{M}>0$.

Proof.

$$\begin{aligned}\sum_{n_1,\ldots,n_{p-1}\in\mathbb{Z}} &\tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\ &\subseteq \sum_{|n_1|,\ldots,|n_{p-1}|\le\overline{M}} \tilde{a}_{n_1}\cdots\tilde{a}_{n_{p-1}}\tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\ &\quad + \sum_{i=1}^{p-1}\max_{a_n\in\tilde{a}_n}\left\{\sum_{n_j\in\mathbb{Z},|n_i|>\overline{M}} |a_{n_1}|\ldots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}|\right\}[-1,1].\end{aligned}$$

Note that

$$\sum_{n_j\in\mathbb{Z},|n_i|>\overline{M}} |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}| = \sum_{n_j\in\mathbb{Z},|n_1|>\overline{M}} |a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}|$$

for $i,\ldots,p-1$.

By Lemma 5.8,

$$\begin{aligned}\sum_{n_j\in\mathbb{Z},|n_1|>\overline{M}} &|a_{n_1}|\cdots|a_{n_{p-1}}||a_{k-(n_1+\cdots+n_{p-1})}| \\ &\le \sum_{n_1>\overline{M}} \frac{A_s}{n_1^s}\frac{\alpha^{p-2}A^{p-1}}{(n_1-k)^s} \\ &\quad + \sum_{n_1<-\overline{M}} \frac{A_s}{(-n_1)^s}\frac{\alpha^{p-2}A^{p-1}}{(k-n_1)^s} \\ &= \alpha^{p-2}A^{p-1}A_s\sum_{n_1>\overline{M}}\frac{1}{n_1^s(n_1-k)^s}\end{aligned}$$

$$
\begin{aligned}
&+ \alpha^{p-2} A^{p-1} A_s \sum_{n_1 > \overline{M}} \frac{1}{n_1^s (k+n_1)^s} \\
&\le \frac{\alpha^{p-2} A^{p-1} A_s}{(\overline{M}-k)^s} \int_{\overline{M}}^{\infty} x^{-s} dx \\
&\quad + \frac{\alpha^{p-2} A^{p-1} A_s}{(\overline{M}+k)^s} \int_{\overline{M}}^{\infty} x^{-s} dx \\
&= \frac{\alpha^{p-2} A^{p-1} A_s}{\overline{M}^{s-1}(s-1)} \left[\frac{1}{(\overline{M}-k)^s} + \frac{1}{(\overline{M}+k)^s} \right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{n_1, \ldots, n_{p-1} \in \mathbb{Z}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\qquad \subseteq \sum_{|n_1|, \ldots, |n_{p-1}| \le \overline{M}} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_1+\cdots+n_{p-1})} \\
&\qquad\quad + \frac{(p-1)\alpha^{p-2} A^{p-1} A_s}{\overline{M}^{s-1}(s-1)} \left[\frac{1}{(\overline{M}-k)^s} + \frac{1}{(\overline{M}+k)^s} \right] [-1,1]. \qquad \blacksquare
\end{aligned}
$$

Corollary 5.11. *For $0 \le k < \overline{M}$,*

$$
\begin{aligned}
&\sum_{n_0, \ldots, n_{p-1} \in \mathbb{Z}} \tilde{c}^p_{n_0} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
&\qquad \subseteq \sum_{|n_0|, \ldots, |n_{p-1}| \le \overline{M}} \tilde{c}^p_{n_0} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
&\qquad\quad + \frac{\alpha^{p-1} A^p C}{\overline{M}^{s-1}(s-1)} \left[\frac{1}{(\overline{M}-k)^s} + \frac{1}{(\overline{M}+k)^s} \right] [-1,1] \\
&\qquad\quad + \frac{(p-1)\alpha^{p-1} A^{p-1} C A_s}{\overline{M}^{s-1}(s-1)} \left[\frac{1}{(\overline{M}-k)^s} + \frac{1}{(\overline{M}+k)^s} \right] [-1,1]
\end{aligned}
$$

for any $\overline{M} > 0$.

Since the sum in these last estimates is finite, we may decide on a case-by-case basis which of these terms contains only variables (a_k with $0 \le k < m$). These terms are contained in $f^{(m)}$ and should not be included when computing bounds for $f^{(m+)}$. We may also compute a better bound of the remaining terms in the sum by using explicit interval bounds for a_n, $n < \overline{M}$, instead of the extended asymptotic bounds.

Corollary 5.12. *For $0 \le k < m$, the error in the kth coordinate map from the neglected higher modes,*

$$
\begin{aligned}
[f^{(m+)}(Z)]_k \subset &\sum_{|n_0|, \ldots, |n_{p-1}| \le \overline{M}} \overline{c^p_{n_0} a_{n_1} \cdots a_{n_{p-1}} a_{k-(n_0+\cdots+n_{p-1})}} \\
&\quad + (A + (p-1)A_s) \frac{(\alpha A)^{p-1} C}{\overline{M}^{s-1}(s-1)} \left[\frac{1}{(\overline{M}-k)^s} + \frac{1}{(\overline{M}+k)^s} \right] [-1,1],
\end{aligned}
$$

where $\overline{c_{n_0} a_{n_1} \cdots a_{n_{p-1}} a_{k-(n_0+\cdots+n_{p-1})}}$ *is* 0 *if all of the indices have absolute value less than* m *and is the interval bound* $\tilde{c}^p_{n_0} \tilde{a}_{n_1} \cdots \tilde{a}_{n_{p-1}} \tilde{a}_{k-(n_0+\cdots+n_{p-1})}$ *otherwise.*

The computations in this section may be used to calculate error bounds (Corollary 5.12) to be used in the construction of the multivalued map and to update explicit interval bounds (Corrollary 5.11). In the following lemma, we use Corollary 5.9 to update the bounds $\tilde{a}_k$, $k > M$.

Lemma 5.13. *Suppose the nonlinearity is* $g(a) = \sum_{p=0}^d c_p a^p$ *with the expansions of the coefficient functions satisfying the decay rules* $c^p_n \in \tilde{c}^p_n := \frac{C_p}{|n|^s}[-1,1]$ *for some constants* C_p*. Then for* $k \geq M$*, we may set new bounds* $\tilde{a}_k$ *to be* $\frac{A_{s+b}}{k^{(s+b)}}[-1,1]$*, where*

$$A_{s+b} = B(C_0 + C_1 A\alpha + \cdots + C_d A^d \alpha^d).$$

Proof. For a fixed $k \geq M$, we have bounds for the projection onto the kth mode of the image of Z under the full map using the bounds for $|b_k|$ and the bounds for the sum given by Corollary 5.9.

$$\begin{aligned}
[F_m(Z)]_k &= b_k \sum_{p=0}^{d} \sum_{n_0,\ldots,n_{p-1}} \tilde{c}^p_{n_0} \tilde{a}_{n_1} \cdots \tilde{a}_{k-(n_0+\cdots+n_{p-1})} \\
&\subseteq \frac{B}{k^b} \frac{1}{k^s} (C_0 + \alpha A C_1 + \cdots + \alpha^d A^d C_d)[-1,1] \\
&= \frac{A_{s+b}}{k^{(s+b)}}[-1,1],
\end{aligned}$$

where A_{s+b} is as given in the statement of the lemma. We preserve isolation by setting the new bounds to be the bounds on the image. ■

6. Example computations.

In this section we apply the previously described methodology for the analysis of the Kot–Schaffer map. We perform rigorous computations in order to prove the existence of (and compute an approximation to) the following three types of invariant sets for Φ:

1. A heteroclinic orbit connecting a neighborhood of a fixed point to a neighborhood of a period two orbit (6.1).
2. An invariant set with chaotic dynamics, i.e., with positive entropy (6.2).
3. A second complicated invariant set, which is described in terms of chaotic symbolic dynamics (6.3).

The computations have been performed using the GAIO [3] and CHomP [17] packages. There are scripts available that will perform the following procedures.

6.1. Connecting orbits.

In this section we show the existence of a heteroclinic orbit of Φ connecting a (small) neighborhood of a fixed point to a (small) neighborhood of a period two point and compute an approximation of this orbit to an accuracy on the order of 10^{-12}. The parameter values for the equivalent countable system (1.4) are as follows:

$$\mu = 3.5, \quad b_k = 2^{-k}, \quad c_0 = 0.8, \quad c_1 = -0.2, \quad \text{and} \quad c_k = 0 \quad \text{for } k > 1. \tag{6.1}$$

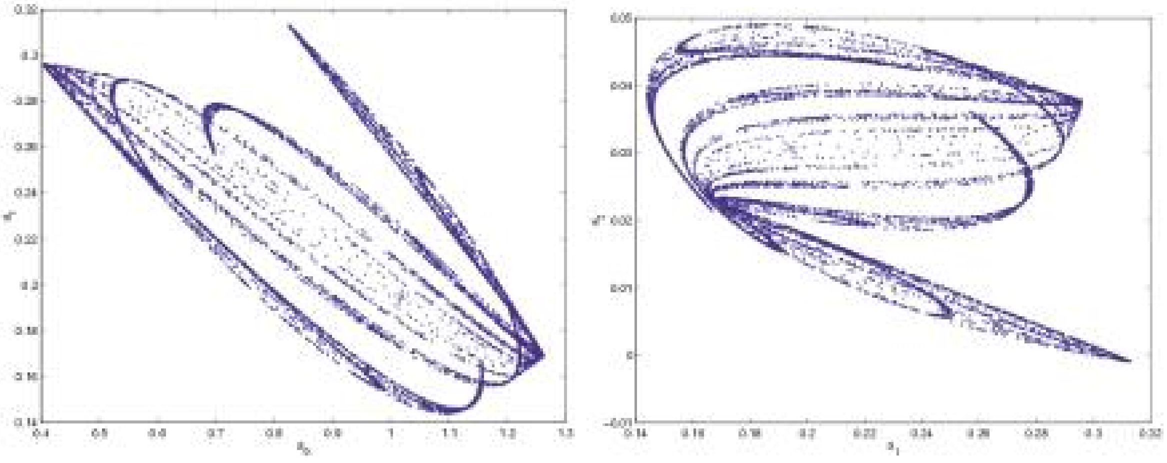


Figure 6.1. *Projections of the set A_K.*

In order to initialize the a priori bounds $a_k^{\pm}$, we run a simulation of $f^{(L)} : \mathbb{R}^L \to \mathbb{R}^L$ for $L = 50$ with the initial condition $A = \{(2^{-k})_{k=0}^{49}\}$, $K_0 = 10$, and $K_1 = 10000$. Figure 6.1 shows two projections of the resulting set A_K. We choose $M = 50$ and an exponential estimate for $k \geq M$ with $A_s = 1$ and $s = 2$ as well as the initial bounds as stated in Table 6.1.

Table 6.1
Initial guess for error bounds $a_k^{\pm}$.

k	a_k^-	a_k^+
0	0.2	1.5
1	0.05	0.5
2	-0.001	0.1
$2 < k < M$	-2^{-k}	2^{-k}

The next step is to choose the (initial) projection dimension m. To this end we compute bounds for the errors $\varepsilon_k^{(m+)}$, $k = 0, \ldots, m-1$, for various m. Based on these values (cf. Table 6.2) we decide to start off with $m = 5$. Again, this is a preliminary choice and at this point in the procedure we cannot guarantee that they will be sufficient later when trying to compute isolating neighborhoods.

Table 6.2
Errors induced by neglecting the higher order modes for different projection dimensions.

m	2	3	4	5	6
$\varepsilon_k^{(m+)}$	0.32139 0.27164	0.18083 0.13052 0.19013	0.058333 0.040833 0.032630 0.047533	0.014583 0.011849 0.010208 0.008158 0.011883	0.0036458 0.0029622 0.0029622 0.0025521 0.0020394 0.0029708

We are now able to compute guesses for invariant sets of $\mathcal{F}^{(5)}$. We do this by running Algorithm 3 with selection criterion (4.1), yielding an approximate covering of the maximal invariant set in the chosen region $W^{(5)} = \prod_{k=0}^{4}[a_k^-, a_k^+]$. In the course of the algorithm the coordinate direction j in which the boxes are subdivided varies cyclically with the number of subdivisions k; i.e., $j = k \bmod 5$. For computing the multivalued map $\mathcal{F}^{(5)}$ on the collections $\mathcal{B}_k$ we employ the heuristic method of mapping test points—here we use a set of 100 points per box distributed randomly according to a uniform distribution. Figure 6.2 shows two projections of the resulting box collection in $\mathbb{R}^5$ after 35 steps of the algorithm.

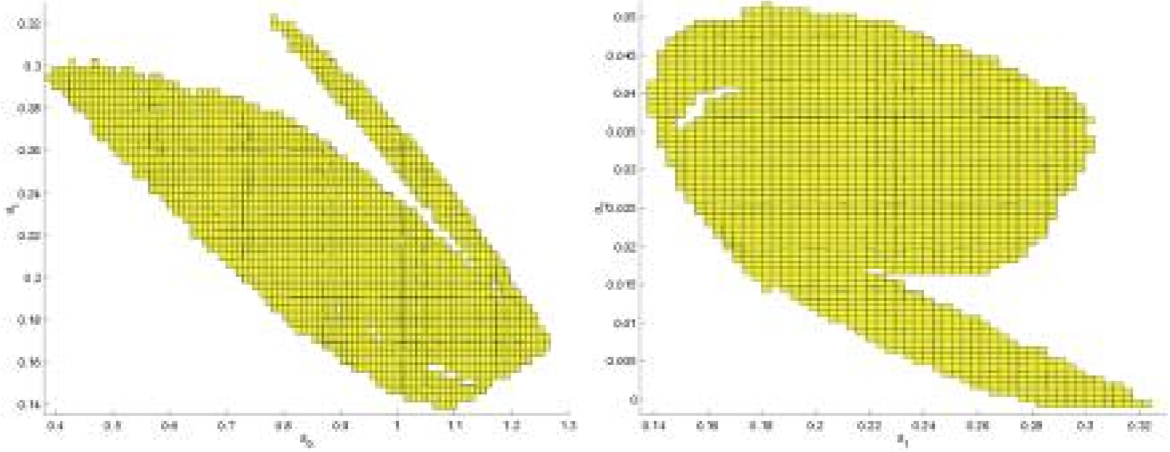
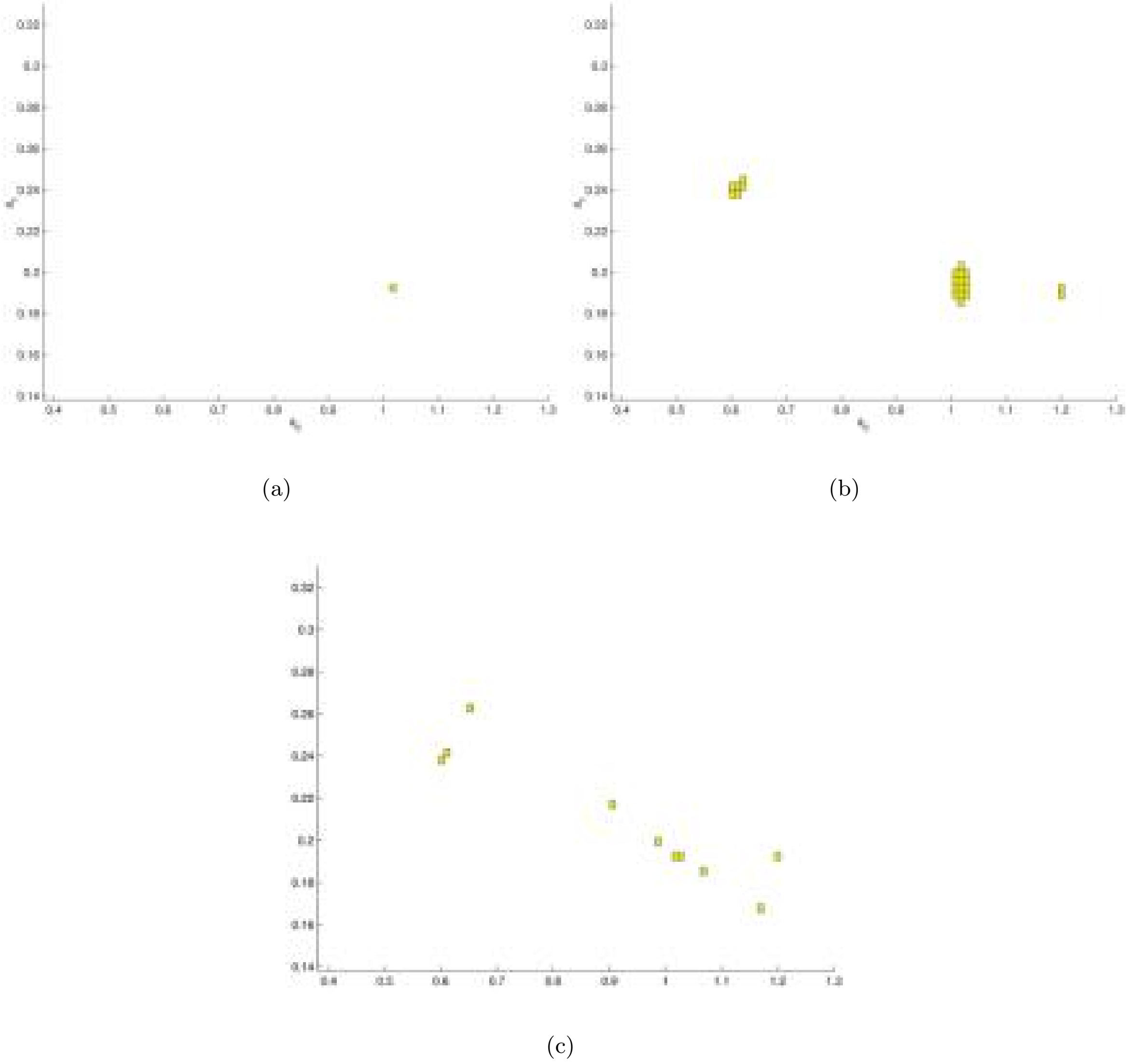


Figure 6.2. *Projections of the (approximate) box covering $\mathcal{B}_{35}$ of the maximal invariant set of $F^{(5)}$ in $W^{(5)}$.*

Before we proceed with extracting the desired dynamical objects from the collection $\mathcal{B}_{35}$ we update the a priori bounds $a_k^{\pm}$, $k = 5, \ldots, 49$, (see section 4.3.2) in order to reduce the error $\varepsilon^{(m+)}$. The resulting $\varepsilon^{(m+)}$ using the updated bounds is smaller than $(0.05, 0.1, 0.2, 0.6, 1.25)^T \cdot 10^{-4}$. Note that this error (rather than the initial error as shown in Table 6.2) determines whether we will be able to perform a useful computation later on. It essentially sets a lower bound to the size of the boxes used to cover the objects of interest. In our case, $\varepsilon^{(m+)}$ is roughly (at least) four times smaller than the radius of the boxes in $\mathcal{B}_{35}$.

Recall that our aim is to compute an orbit connecting a fixed point to a period two point of Φ. Guesses for isolating neighborhoods of the periodic points of $f^{(5)}$ are easily obtained by considering the periodic points of $\mathcal{F}^{(5)} : \mathcal{B}_{35} \rightrightarrows \mathcal{B}_{35}$ as laid out in section 4.1. Figure 6.3 shows the guess $\tilde{\mathcal{I}}_1$ for the fixed points and the guess $\tilde{\mathcal{I}}_2$ for the period two points. Finally, we obtain a guess $\tilde{\mathcal{I}}_c$ for a connecting orbit by computing the shortest path from the box in $\tilde{\mathcal{I}}_1$ to some box in $\tilde{\mathcal{I}}_2$ (cf. Figure 6.3(c); see again section 4.1). Using the collection $\tilde{\mathcal{I}}_c$ as input we now intend to use Algorithm 2 to compute an isolating neighborhood for the multivalued map $F^{(5)}$. To this end we need to deal with a (true) enclosure of $F^{(5)}$ (in contrast to the heuristic one we used so far in order to obtain our initial guesses). Bounds on the errors $\varepsilon^{(m+)}$ are computed using the results of section 5.2 and updated using the procedures discussed in section 4.3.2.

By running Algorithm 2 with input $\tilde{\mathcal{I}}_c$, now we obtain a combinatorial isolating neighborhood $\mathcal{I}$ for $\mathcal{F}^{(5)}$ (cf. Figure 6.4). The union $|\mathcal{I}|$ of these boxes is an isolating neighborhood



(a)

(b)

(c)

Figure 6.3. *Guesses for isolating neighborhoods of* (a) *the fixed points,* (b) *the period two points (note that the fixed points are covered, too), and* (c) *a connecting orbit of* $f^{(5)}$ *(projections onto the first two coordinates).*

for $F^{(5)}$ (cf. [22]). Using (1.13) we finally compute a corresponding index pair $(|\mathcal{N}_1|,|\mathcal{N}_0|)$. Figure 6.5 shows two different views of these sets, where the exit set $|\mathcal{N}_0|$ corresponds to the red boxes. The resulting relative homology groups of the pair $N = (|\mathcal{N}_1|, |\mathcal{N}_0|)$ are

$$H_*(|\mathcal{N}_1|, |\mathcal{N}_0|) \cong (0,\ \mathbb{Z}^7,\ 0,\ 0,\ \ldots).$$

Obviously, the associated homology map $f^{(5)}_{N*} : H_*(|\mathcal{N}_1|, |\mathcal{N}_0|) \to H_*(|\mathcal{N}_1|, |\mathcal{N}_0|)$ induced by $f^{(5)}_N$ is trivial everywhere but level one. For an appropriate choice of basis, $f^{(5)}_{N,1}$ can be

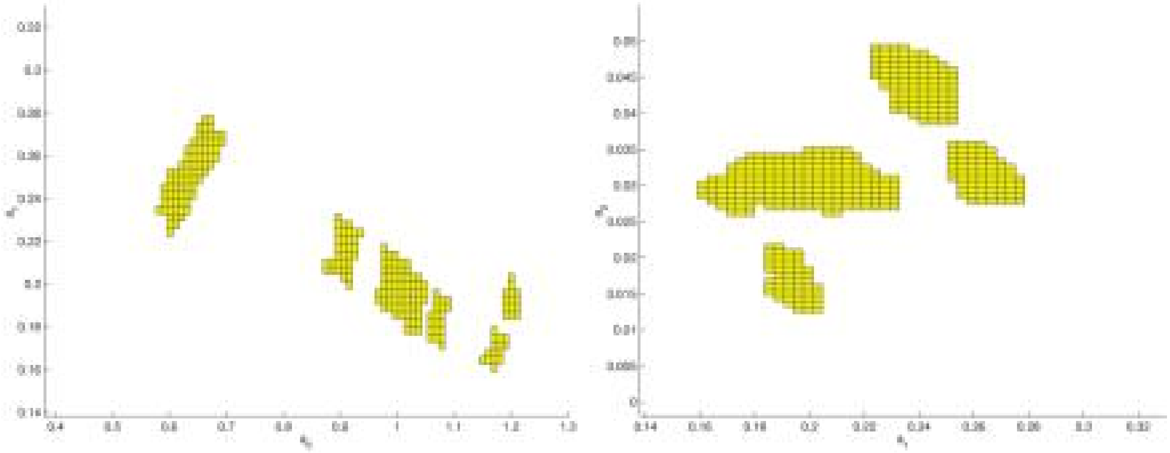


Figure 6.4. *Combinatorial isolating neighborhood for $\mathcal{F}^{(5)}$ (projections onto the first and the second two coordinates, respectively).*

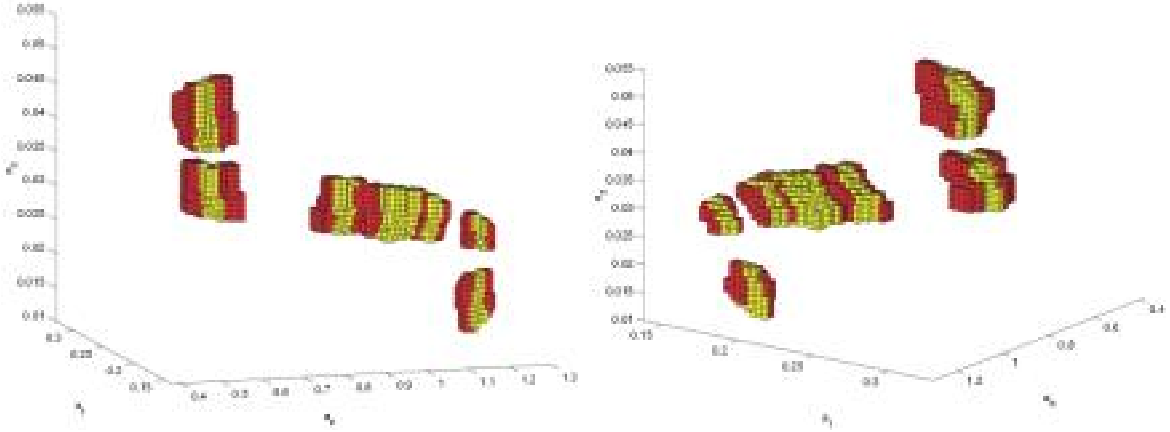


Figure 6.5. *Index pair for the multivalued map $F^{(5)}$ (projections onto the first three coordinates). The red boxes correspond to the exit set $|\mathcal{N}_0|$.*

expressed as the following matrix:

$$f_{N,1}^{(5)} := \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

An important observation is that each generator of $H_1(|\mathcal{N}_1|, |\mathcal{N}_0|)$ lies in a distinct connected component of $|\mathcal{N}_1| \setminus |\mathcal{N}_0|$. Using the same notation as in section 2.2 we label these components as $B_1, \dots, B_7$. The multivalued map $F^{(5)}$ provides us with the following information concerning the dynamics between the boxes:

1. $F^{(5)}(B_1) \cap B_j \neq \emptyset$ only if $j = 1, 2$.
2. For $i = 2, \dots, 6$, $F^{(5)}(B_i) \cap B_j \neq \emptyset$ only if $j = i + 1$.
3. $F^{(5)}(B_7) \cap B_j \neq \emptyset$ only if $j = 6$.

In particular, we can consider the following index maps, obtained from $f^{(5)}_{N,1}$ by projection onto the corresponding connected components:

$$f^{(5)}_{B_1,1} := \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad f^{(5)}_{B_6 \cup B_7,1} := \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{bmatrix}.$$

Observe that $\operatorname{tr} f^{(5)}_{B_1} = -1$ and $\operatorname{tr} (f^{(5)}_{B_6 \cup B_7})^2 = -2$.

There are several conclusions that can be drawn from this information. First, $\operatorname{Inv}(B_1, f^{(5)}) \neq \emptyset$ and $\operatorname{Inv}(B_6 \cup B_7, f^{(5)}) \neq \emptyset$. In fact, by (2.5) we know that $\operatorname{Inv}(B_1, f^{(5)})$ contains a fixed point and $\operatorname{Inv}(B_6 \cup B_7, f^{(5)})$ contains a periodic point. The second and third pieces of information concerning the multivalued map $F^{(5)}$ indicate that this periodic orbit has minimal period 2.

Finally, $f^{(5)}_{N,1}$ is not shift equivalent to $f^{(5)}_{B_1,1} \oplus f^{(5)}_{B_6 \cup B_7,1}$; thus the Conley indices of $\operatorname{Inv}(I, f^{(5)})$ and $\operatorname{Inv}(B_1 \cup B_6 \cup B_7, f^{(5)})$ are different. This in turn implies that $\operatorname{Inv}(I, f^{(5)}) \neq \operatorname{Inv}(B_1 \cup B_6 \cup B_7, f^{(5)})$. Again, returning to the information concerning the multivalued map $F^{(5)}$ we can conclude that $\operatorname{Inv}(I, f^{(5)})$ must contain an orbit whose omega limit set is contained in $B_6 \cup B_7$ and whose alpha limit set is contained in B_1.

Since these results concerning the existence of orbits follow from the algebra of the Conley index, after checking condition (2.4) we can immediately carry over this existence result to the infinite-dimensional system (1.4), and thus to (1.1).

However, we aim for a more precise localization of the detected objects and continue with tightening the computed covering as described in section 4.3. We end up with a collection $\mathcal{I}^{(11)}$ of 24,078 boxes in $\mathbb{R}^{11}$ and—employing Corollaries 2.4 and 2.5—the following theorem.

Theorem 6.1. *For the parameter values* (6.1) *the map* Φ *possesses the following orbits, all of which lie in the set*

$$|\mathcal{I}^{(11)}| \times \prod_{k=11}^{49} [a_k^-, a_k^+] \times \prod_{k=50}^{\infty} \frac{1}{2^k}[-1, 1],$$

where the $a_k^\pm$ *are the final bounds:*

1. *A fixed point* p *with the property that*

$$\|p - p^*\|_{L^2} < 3 \cdot 10^{-12}, \quad \|p - p^*\|_{C^0} < 3.6 \cdot 10^{-12},$$

where the Fourier coefficients of p^* *(given in Table* 6.3*) are determined by* B_1*.*

2. *A periodic orbit* $q = \{q_1, q_2\}$ *with the property that*

$$\|q_i - q_i^*\|_{L^2} < 3.3 \cdot 10^{-12}, \quad \|q_i - q_i^*\|_{C^0} < 4.3 \cdot 10^{-12}, \qquad i = 1, 2,$$

where the Fourier coefficients of q_1^* *and* q_2^* *(given in Table 6.4) are determined by* B_6 *and* B_7*, respectively.*

3. *A heteroclinic orbit* $a = (a_j)_{j \in \mathbb{Z}}$*, with the property that*

$$\delta(\alpha(a), p) < 3.5 \cdot 10^{-12}, \quad \delta(\omega(a), q) < 3.8 \cdot 10^{-12},$$

where $\delta(\cdot, \cdot)$ *is the distance in the Hausdorff metric on compact sets.*

Table 6.3

Fourier coefficients of p^**.*

k	$(p^*)_k$
0	1.01701222469896
1	0.194337336695483
2	0.030518985313998
3	0.00388416812157288
4	0.000406689870061427
5	$3.59686642677538 \cdot 10^{-5}$
6	$2.75312512992254 \cdot 10^{-6}$
7	$1.85783274541004 \cdot 10^{-7}$
8	$1.12063486092261 \cdot 10^{-8}$
9	$6.10776632026745 \cdot 10^{-10}$
10	$3.03735627584813 \cdot 10^{-11}$
≥ 11	0

Table 6.4

Fourier coefficients of q_1^* *and* q_2^**.*

k	$(q_1^*)_k$	$(q_2^*)_k$
0	0.612404314978377	1.20068110795964
1	0.240548407216622	0.191505924758062
2	0.0427803169880592	0.0186218996298865
3	0.00371012084664812	0.00293452057215404
4	0.000286872331997756	0.000493085588431086
5	$2.85759422150114 \cdot 10^{-5}$	$4.83837782923445 \cdot 10^{-5}$
6	$2.46114548652796 \cdot 10^{-6}$	$3.16438225821576 \cdot 10^{-6}$
7	$1.61009990418761 \cdot 10^{-7}$	$1.78892095960431 \cdot 10^{-7}$
8	$9.62062994427963 \cdot 10^{-9}$	$1.00726290466515 \cdot 10^{-8}$
9	$5.95120614532175 \cdot 10^{-10}$	$5.28784988682338 \cdot 10^{-10}$
10	$3.43628779469884 \cdot 10^{-11}$	$2.46555443216212 \cdot 10^{-11}$
≥ 11	0	0

6.2. Positive entropy. As a second example we now reveal the presence of complicated dynamics for the parameter values chosen in the previous example.

Using the methods employed in the previous example, we first construct an isolating neighborhood, which contains a heteroclinic cycle between a fixed point and a period two point of Φ. To this end we compute guesses $\tilde{\mathcal{I}}_{c_{12}} \subset \mathcal{B}_{36}$ and $\tilde{\mathcal{I}}_{c_{21}} \subset \mathcal{B}_{36}$ for orbits connecting the fixed point to a period two point of $\mathcal{F}^{(5)}$ and vice versa as described in the previous example. Then we use Algorithm 2 with input $\tilde{\mathcal{I}}_{c_{12}} \cup \tilde{\mathcal{I}}_{c_{21}}$ to construct an isolating neighborhood $\mathcal{I}^{(5)}$ for $\mathcal{F}^{(5)}$. A corresponding index pair $(\mathcal{N}_1, \mathcal{N}_0)$ is shown in Figure 6.6.

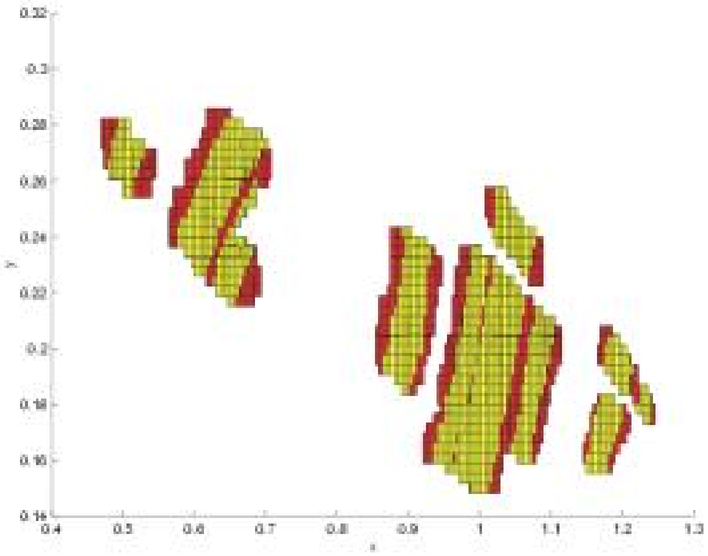


Figure 6.6. *Index pair for $\mathcal{F}^{(5)}$ (projection onto the first two coordinates). The yellow boxes cover the invariant set; the red boxes constitute the exit set.*

The relative homology of the index pair $N = (|\mathcal{N}_1|, |\mathcal{N}_0|)$ for $f^{(5)}$ is

$$H_*(|\mathcal{N}_1|, |\mathcal{N}_0|) \cong (0,\ \mathbb{Z}^{14},\ 0,\ 0,\ \ldots)$$

and the map $f^{(5)}_{N*} : H_*(|\mathcal{N}_1|, |\mathcal{N}_0|) \to H_*(|\mathcal{N}_1|, |\mathcal{N}_0|)$ induced by $f^{(5)}$ in homology is trivial at all levels but level one, on which $f^{(5)}_{N*}$ acts on the 14 ordered generators as follows:

$$f^{(5)}_{N,1} = \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&1&0&1&0\\ 0&0&0&0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&1&1&0&0&0&1&0&0\\ 0&0&0&1&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0\\ 1&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}.$$

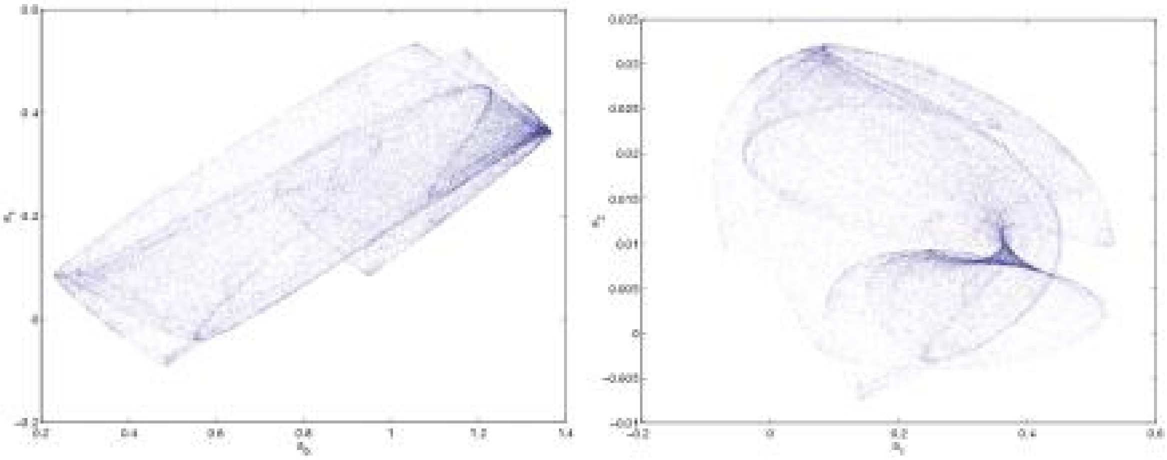


Figure 6.7. *Projections of the set A_K.*

The spectral radius of $f_{N,1}^{(5)}$ is bigger than 1.2, therefore—according to [2] and using Corollary 2.5—we have the following theorem.

Theorem 6.2. *For the values given in (6.1), there exists an invariant set contained in*

$$|\mathcal{I}^{(5)}| \times \prod_{k=6}^{49} [a_k^-, a_k^+] \times \prod_{k=50}^{\infty} \frac{1}{2^k}[-1,1], \quad j \in \mathbb{Z},$$

(where $[a_k^-, a_k^+]$ are the final bounds), on which Φ exhibits positive entropy.

6.3. A 2d-unstable horseshoe. As a last example we prove the existence of and localize another invariant set on which Φ exhibits complicated dynamics. In contrast to the previous computation, this time we aim for the dimension of its unstable manifold to be two.

We are considering the map (1.4) with the following values for its parameters:

$$\begin{aligned} &\mu = 3.8, \quad b_0 = 1, \quad b_1 = 0.99, \quad b_2 = 0.1, \quad b_3 = 0.025, \quad b_k = 2^{-k}, \ k \geq 3, \\ &c_0 = 0.8, \quad c_1 = -0.2, \quad \text{and} \quad c_k = 0 \text{ for } k > 1. \end{aligned}$$

Again, we start by running a simulation of $f^{(L)} : \mathbb{R}^L \to \mathbb{R}^L$ for $L = 100$ with the initial point $A = \{(10^{-4}, 0, \ldots, 0)\}$, $K_0 = 100$, and $K_1 = 50000$. Figure 6.7 shows two projections of the resulting set A_K. We choose $m = 6$, $M = 50$, and an exponential estimate for $k \geq M$ with $A_s = 1$ and $s = 2$ as well as the initial bounds as indicated in Table 6.5.

Again, in order to compute guesses for invariant sets we first compute a covering of the maximal invariant set of $F^{(6)}$ by running 44 steps of Algorithm 3 with selection criterion (4.1). We employ the heuristic method of mapping 200 randomly distributed (according to a uniform distribution) test points per box in order to compute $\mathcal{F}^{(6)}$. The resulting collection consists of 1,130,128 boxes. As in the previous examples, we now update the bounds. The updated bounds lead to an error $\varepsilon^{(m+)}$, which is smaller than $(2 \cdot 10^{-7}, 5 \cdot 10^{-7}, 2 \cdot 10^{-7}, 2 \cdot 10^{-6}, 7 \cdot 10^{-6}, 2 \cdot 10^{-5})^T$. Similarly to the previous example, we compute guesses $\tilde{\mathcal{I}}_{c_{12}} \subset \mathcal{B}_{44}$ and $\tilde{\mathcal{I}}_{c_{21}} \subset \mathcal{B}_{44}$ for orbits connecting the fixed point to a period two point of $F^{(5)}$ and vice versa.

Table 6.5
Initial guess for error bounds $a_k^\pm$.

k	a_k^-	a_k^+
0	0.1601	1.7781
1	-0.06131	0.70231
2	-0.0036184	0.04173
3	-0.00032171	0.033337
$3 < k < M$	-2^{-k}	2^{-k}

Again, we use Algorithm 2 with input $\tilde{\mathcal{I}}_{c_{12}} \cup \tilde{\mathcal{I}}_{c_{21}}$ to construct an isolating neighborhood $\mathcal{I}^{(6)}$ for $\mathcal{F}^{(6)}$. The relative homology of the corresponding index pair $N = (|\mathcal{N}_1|, |\mathcal{N}_0|)$ for $f^{(6)}$ is

$$H_*(|\mathcal{N}_1|, |\mathcal{N}_0|) \cong (0, \mathbb{Z},\ \mathbb{Z}^{18},\ 0,\ 0,\ \dots).$$

The neighborhood $|\mathcal{N}_1|$ consists of 11 connected components, $A, B, \dots, K$. We label the 18 generators of $H_2(|\mathcal{N}_1|, |\mathcal{N}_0|)$ according to their location in these components as shown in Figure 6.8. The combinatorial map $\mathcal{F}^{(6)}$ on the boxes induces a map on the components $A, B, \dots, K$. This map is shown in Figure 6.8(a), together with the homology map on $H_2(|\mathcal{N}_1|, |\mathcal{N}_0|)$ (Figure 6.8(b)).

We now use the index information to prove that there exists a set contained in $|\mathcal{N}_1| \times \Pi_{k\geq 6}[a_k^-, a_k^+]$ on which the full map Φ exhibits symbolic dynamics as depicted by the transition graph T_1 shown in Figure 6.8(a).

Let $\Sigma = \{(Z_i)_{i\in\mathbb{Z}} \mid Z_i \in \{A, B, \dots, K\}\}$ and let $\sigma : \Sigma \to \Sigma$ be the shift map. Consider the subset $\Sigma_* = \{(Z_i)_i \in \Sigma \mid (Z_i, Z_{i-1}) \text{ is an edge in } T_1\}$. As in section 2, define $\rho : |\mathcal{N}_1| \to \Sigma_*$ by $\rho(x) = (Z_i)_{i\in\mathbb{Z}}$, where Z_i is the connected component of $|\mathcal{N}_1|$ containing $(f^{(6)})^i(x), i \in \mathbb{Z}$. A result of [20] allows us to decompose the index information into maps on the connected components of $|\mathcal{N}_1|$. Let $f^{(6)}_{Z,2}$, $Z \in \{A, \dots, K\}$, be the index map obtained from $f^{(6)}_{N,2}$ by projecting onto the connected component Z. According to Szymczak in [23], for every n there exists at least one trajectory under $f^{(6)}$ through $Z_1, \dots, Z_n$ provided that the composition $f^{(6)}_{Z_1,2} \circ \cdots \circ f^{(6)}_{Z_n,2}$ is not nilpotent. In the limit, any infinite sequence of symbols in Σ_* corresponds to at least one trajectory under $f^{(6)}$ through the given components. Furthermore, if the Lefschetz number of $f^{(6)}_{Z_1*} \circ \dots \circ f^{(6)}_{Z_n*}$ is nonzero, then a periodic orbit under $f^{(6)}$ through components $Z_1, \dots, Z_n$ exists (see also (2.5)). Since trace $f^{(6)}_{N,i} = 0$ for $i \neq 2$ (we have $f^{(6)}_{N,1} = 0$), the Lefschetz number is nonzero provided that the trace of the composition of restricted index maps for $H_2(|\mathcal{N}_1|, |\mathcal{N}_0|)$ is nonzero. This can be shown for the composition of restricted index maps for any periodic sequence given by the transition graph in Figure 6.8. By extension, ρ maps onto Σ_*, the closure of periodic orbits given in the transition graph T_1 shown in Figure 6.8(a).

By construction, ρ is a semiconjugacy between $f^{(6)}$ on the (nonempty) invariant set contained in $|\mathcal{N}_1| \setminus |\mathcal{N}_0|$ and the shift map σ on Σ_*. After checking that the lifting condition (2.4) is satisfied, we have the following theorem.

Theorem 6.3.*For the parameter values* (6.2) *there is an invariant set contained in* $|\mathcal{N}_1| \times \Pi_{k\geq 6}[a_k^-, a_k^+]$ *(where* $[a_k^-, a_k^+]$ *are the final bounds), on which* Φ *is semiconjugate to the subshift given by the transition graph* T_1 *shown in Figure* 6.8(a).

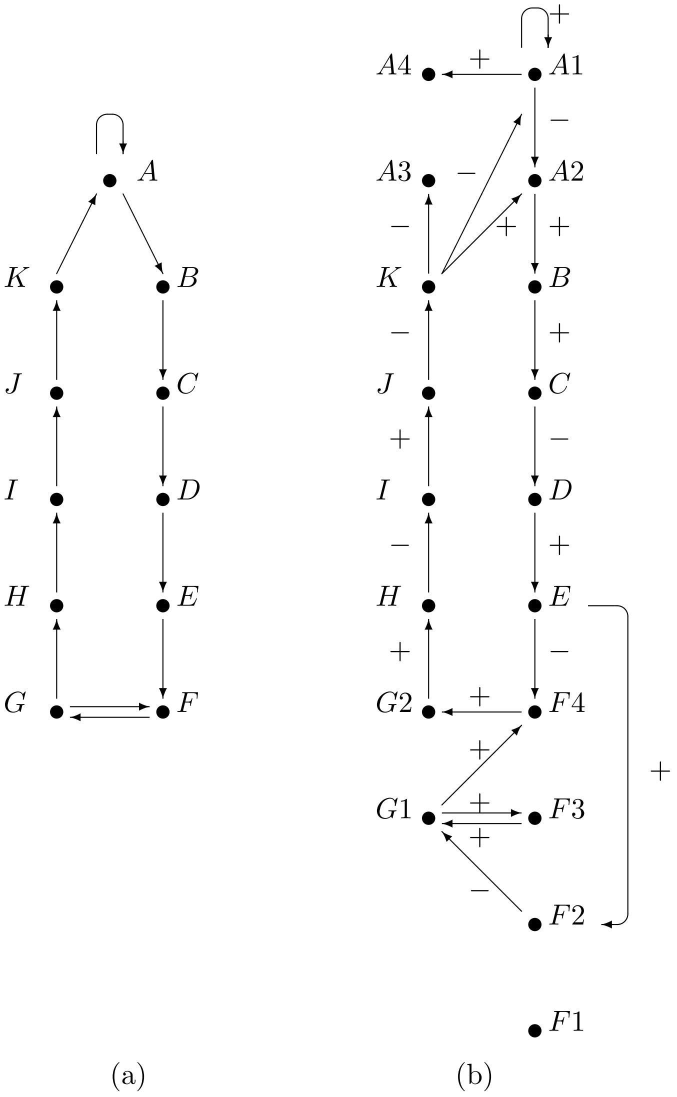


Figure 6.8. (a) *The graph T_1. It indicates the existence of* 11 *connected components of the isolating neighborhood. The arrows indicate how the connected components are related by the map $F^{(6)}$.* (b) *The graph T_2. This graph indicates the relative homology classes within each connected component, and the edges indicate how the generators of these homology classes are mapped into each other. The* ± *signs indicate whether the matrix entry for $F_2^{(6)} : H_2(N_1, N_0) \to H_2(N_1, N_0)$ is* ±1.

Acknowledgments. The authors would like to thank Madjid Allili and Paweł Pilarczyk for providing code and assistance in doing the homology computations.

REFERENCES

[1] G. Arioli and P. Zgliczyński, *Symbolic dynamics for the Hénon-Heiles Hamiltonian on the critical level*, J. Differential Equations, 171 (2001), pp. 173–202.

[2] A. W. Baker, *Lower bounds on entropy via the Conley index with application to time series*, Topology Appl., 120 (2002), pp. 333–354.

[3] M. Dellnitz, G. Froyland, and O. Junge, *The algorithms behind GAIO-set oriented numerical methods for dynamical systems*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, Berlin, 2001, pp. 145–174; 805–807.

[4] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.

[5] M. Dellnitz, O. Junge, M. Rumpf, and R. Strzodka, *The computation of an unstable invariant set inside a cylinder containing a knotted flow*, in Proceedings of the International Conference on Differential Equations, World Scientific, River Edge, NJ, 2000, pp. 1053–1059.

[6] E. W. Dijkstra, *A note on two problems in connection with graphs*, Numer. Math., 1 (1959), pp. 269–271.

[7] M. Eidenschink, *Exploring Global Dynamics: A Numerical Algorithm Based on the Conley Index Theory*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1995.

[8] J. Franks and D. Richeson, *Shift equivalence and the Conley index*, Trans. Amer. Math. Soc., 352 (2000), pp. 3305–3322.

[9] J. K. Hale and G. Raugel, *Regularity, determining modes and Galerkin methods*, J. Math. Pure Appl. (9), 82 (2003), pp. 1075–1136.

[10] O. Junge, *Rigorous discretization of subdivision techniques*, in Proceedings of the International Conference on Differential Equations, World Scientific, River Edge, NJ, 2000, pp. 916–918.

[11] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, Appl. Math. Sci. 157, Springer-Verlag, New York, 2004.

[12] T. Kaczynski and M. Mrozek, *Conley index for discrete multivalued dynamical systems*, Topology Appl., 65 (1995), pp. 83–96.

[13] O. Knüppel, *BIAS – Basic Interval Arithmetic Subroutines*, Technische Universität Hamburg–Harburg, Hamburg, Germany, 1993.

[14] O. Knüppel, *PROFIL – Programmer's Runtime Optimized Fast Interval Library*, Technische Universität Hamburg–Harburg, Hamburg, Germany, 1993.

[15] M. Kot and W. M. Schaffer, *Discrete-time growth-dispersal models*, Math. Biosci., 80 (1986), pp. 109–136.

[16] K. Mischaikow and M. Mrozek, *Conley index*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 393–460.

[17] K. Mischaikow, M. Mrozek, and P. Pilarczyk, *Graph Approach to the Computation of the Homology of Continuous Maps*, Preprint, 2004.

[18] K. Mischaikow, M. Mrozek, and A. Szymczak, *Chaos in the Lorenz equations: A computer assisted proof.* III. *Classical parameter values*, J. Differential Equations, 169 (2001), pp. 17–56.

[19] M. Mrozek, *Shape index and other indices of Conley type for local maps on locally compact Hausdorff spaces*, Fund. Math., 145 (1994), pp. 15–37.

[20] A. Szymczak, *The Conley index for decompositions of isolated invariant sets*, Fund. Math., 148 (1995), pp. 71–90.

[21] A. Szymczak, *The Conley index and symbolic dynamics*, Topology, 35 (1996), pp. 287–299.

[22] A. Szymczak, *A combinatorial procedure for finding isolating neighbourhoods and index pairs*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1075–1088.

[23] A. Szymczak, *Index Pairs: From Dynamics to Combinatorics and Back*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1999.

[24] W. Tucker, *A rigorous ODE solver and Smale's* 14*th problem*, Found. Comput. Math., 2 (2002), pp. 53–117.

[25] J. Zemke, b4m, *A Free Interval Arithmetic Toolbox for Matlab Based on BIAS*, Technische Universität Hamburg–Harburg, Hamburg, Germany, 1998, http://www.ti3.tu-harburg.de/zemke/b4m.

[26] P. Zgliczyński and K. Mischaikow, *Rigorous numerics for partial differential equations: The Kuramoto-Sivashinsky equation*, Found. Comput. Math., 1 (2001), pp. 255–288.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 2, pp. 161–190

Computing One-Dimensional Stable Manifolds and Stable Sets of Planar Maps without the Inverse*

J. P. England†, B. Krauskopf†, and H. M. Osinga†

Abstract. We present an algorithm to compute the one-dimensional stable manifold of a saddle point for a planar map. In contrast to current standard techniques, here it is not necessary to know the inverse or approximate it, for example, by using Newton's method. Rather than using the inverse, the manifold is grown starting from the linear eigenspace near the saddle point by adding a point that maps back onto an earlier segment of the stable manifold. The performance of the algorithm is compared to other methods using an example in which the inverse map is known explicitly. The strength of our method is illustrated with examples of noninvertible maps, where the stable set may consist of many different pieces, and with a piecewise-smooth model of an interrupted cutting process. The algorithm has been implemented for use in the DsTool environment and is available for download with this paper.

Key words. discrete-time system, planar map, stable manifold, noninvertible map, stable set, piecewise-smooth map

AMS subject classifications. 37D10, 37M20, 65P30

DOI. 10.1137/030600131

1. Introduction.

Many interesting dynamical systems arising in applications can be described by maps; examples are mechanical systems, laser systems, electronic circuits, biological processes, and interactions between populations. An important class consists of Poincaré maps of continuous vector fields, including, in particular, periodically driven systems such as the driven damped pendulum or the forced Van der Pol oscillator; see [9, 31] for general references. Poincaré maps are also often used when the periodicity is intrinsic in the vector field. For such systems, however, the Poincaré map may not be defined globally, and thus the resulting discrete system is not a global diffeomorphism. Note that there is generally no explicit expression for the Poincaré map of a given vector field. In a number of applications the map is defined explicitly; such maps include the Ikeda map [12, 15], as well as the example of a highly interrupted cutting process [4, 5, 33] discussed in this paper.

The goal of analyzing a given system is to understand its overall, or global, behavior. To this end, one needs to find special invariant objects. For two-dimensional discrete-time systems, these are the attractors, the repellers, and the saddle points, together with their stable and unstable manifolds. In particular, these manifolds tell us a lot about the dynamics of the system. For example, they may form the boundaries of basins of attraction. Furthermore,

[*] Received by the editors April 15, 2003; accepted for publication (in revised form) by M. Golubitsky January 24, 2004; published electronically May 19, 2004.

http://www.siam.org/journals/siads/3-2/60013.html

[†] Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (james.england@bristol.ac.uk, b.krauskopf@bristol.ac.uk, h.m.osinga@bristol.ac.uk). The research of the first author was supported by grant GR/R94572/01 from the Engineering and Physical Sciences Research Council (EPSRC).

intersections between stable and unstable manifolds may lead to homoclinic or heteroclinic tangle and chaotic dynamics. Stable and unstable manifolds are global objects that cannot normally be found analytically and, hence, must be computed numerically.

Several methods have been developed to compute the stable and unstable manifolds of a saddle fixed point. We concentrate here on the simplest case, which is when these manifolds are one-dimensional. Most algorithms compute the manifold by starting from a local approximation near the saddle point. A common technique is that of iterating a fundamental domain [13, 14, 29, 30, 35]. This involves iterating a local section of the manifold to build up connecting pieces of manifold. An alternate approach, also used here, is to grow the manifold point by point and to parameterize by its arclength [17]. All these methods have in common the fact that the stable manifold is computed as the unstable manifold of the inverse map. This is not a problem when one is dealing with a globally defined Poincaré map. Such a map always has an inverse, which can be computed numerically by following the flow of the vector field backward in time.

However, except in idealized circumstances, an *explicit map* that arises in a particular application does not have an explicit inverse. The expression for such a map is generally quite intricate, contains many parameters, and is often written as the composition of several submaps. Therefore, it is usually impossible to derive an explicit expression for its inverse. What is more, the map may not even be invertible, meaning that there may be several branches of inverses. Consequently, the standard algorithms requiring the inverse cannot be used to compute stable sets (i.e., the generalization of stable manifolds; see section 2.1) of saddle points in this case. If the inverse exists but is not available in closed form, it is possible to approximate it with, say, Newton's method. To this end, a good seed (i.e., initial guess) is required for finding the inverse. Moreover, one needs to know the Jacobian of the map, which may not be well defined or may become singular at certain points. Such requirements are often not practical in applications, as is illustrated in the example of the highly interrupted cutting process [33] in section 4.5.

In this paper we present an algorithm to compute the stable manifold, or the stable set, of a saddle point of a planar map without requiring any knowledge of its inverse map either explicitly or approximately. In particular, our algorithm can also be used in the case where the map is noninvertible so that multiple pre-images may exist. The algorithm grows branches of the stable manifold, or the stable set, point by point. A new point is found by searching for an intersection point of the image of a circle around the last computed point with the part of the manifold that was already computed. We refer to this as the search circle (SC) algorithm. By prespecifying certain accuracy parameters, we achieve a good resolution of the computation.

Another computational technique that does not require knowledge of the inverse map is the proper interior maximum triple (PIM-triple) technique [25], also known as straddling, to compute chaotic saddles. It can be used to find *saddle straddle trajectories* that remain in a given region for a long period of time. If the region contains a saddle point, then points on the stable manifold are found because they remain in the region indefinitely. However, the PIM-triple technique has not been used to compute long pieces of stable manifolds.

The SC algorithm is fully implemented in the DsTool environment [3]. The user may download the code for linking with the Tcl/Tk version of DsTool; see [26] for details. All images in this paper were produced with this implementation.

The performance of the SC algorithm is demonstrated with several examples. First we consider a simple test map, the shear map, for which the inverse is known explicitly [17]. Then we compute the stable manifold of the modified Ikeda map (without an explicit inverse), which was computed in [16] by approximating the inverse with Newton's method. Our third example is a noninvertible map with a stable set that crosses regions where there are two pre-images as well as regions where there are no pre-images. Next, we consider the noninvertible map introduced in [24]. The global stable set for this map contains infinitely many disjoint pieces, and we compute a number of them for the first time. Our last example is a map of a highly interrupted cutting process [4, 5, 33]. It does not have an explicit inverse, and the map is only piecewise smooth.

The outline of this paper is as follows. In section 2 we introduce the notation and explain some of the mathematical concepts. In particular, we give a brief overview of the relevant theory for noninvertible systems in section 2.1. The SC algorithm is explained in section 3 and we investigate its performance with the examples in section 4. In section 5 we draw conclusions and point to open problems.

2. Mathematical setting.

We consider a discrete dynamical system with a two-dimensional phase space, given by a continuous map

$$f : \mathbb{R}^2 \mapsto \mathbb{R}^2.$$

If f is not orientation preserving, we consider its second iterate. We assume that f has a fixed point $x_0 = f(x_0)$ and that f is differentiable in a neighborhood of x_0, but it may not have a unique inverse.

The fixed point x_0 of f is a saddle if the Jacobian matrix $Df(x_0)$ has one stable eigenvalue λ_s inside the unit circle and one unstable eigenvalue λ_u outside the unit circle. The stable manifold theorem [27] guarantees that there exist local stable and unstable manifolds $W^s_{loc}(x_0)$ and $W^u_{loc}(x_0)$ tangent at x_0 to the stable and unstable eigenspaces $E^s(x_0)$ and $E^u(x_0)$, respectively.

The *global unstable manifold* $W^u(x_0)$ of x_0 consists of points that converge to x_0 under backward iteration of the map f. In terms of forward iterates, this is defined as

$$W^u(x_0) = \left\{ x \in \mathbb{R}^2 \;\middle|\; \exists \left\{q_k\right\}_{k=0}^{\infty}, \, q_0 = x \text{ and } f\left(q_{k+1}\right) = q_k, \text{ such that } \lim_{k\to\infty} q_k = x_0 \right\}.$$

The unstable manifold can be expressed in terms of the successive union of images of the local unstable manifold $W^u_{loc}(x_0)$, namely,

$$W^u(x_0) = \bigcup_{n=1}^{\infty} f^n(W^u_{loc}(x_0)).$$

Note that, even if f is noninvertible, the images of $W^u_{loc}(x_0)$ will be unique and, hence, $W^u(x_0)$ is an embedded manifold in the phase space. In particular, we are justified in speaking of $W^u(x_0)$ as the unstable manifold.

The *global stable set* $W^s(x_0)$ of x_0 is defined as the set of points that converge to x_0 under forward iteration of f,

$$W^s(x_0) = \Big\{ x \in \mathbb{R}^2 \mid f^n(x) \to x_0 \text{ as } n \to \infty \Big\}.$$

It can be obtained as the union of the successive pre-images of $W^s_{loc}(x_0)$. If the map f is invertible, then $W^s(x_0)$ is an embedded manifold and one speaks of $W^s(x_0)$ as the stable manifold. However, if multiple inverses exist, then the stable set $W^s(x_0)$ may consist of disjoint pieces. In particular, this set is not an embedded manifold, and this is why one speaks of $W^s(x_0)$ as the global stable set [22]. Throughout this paper, we refer to the unique piece of $W^s(x_0)$ that contains the fixed point x_0 as the *primary manifold*. Because it contains $W^s_{\text{loc}}(x_0)$, the primary manifold is indeed a one-dimensional manifold.

2.1. Background on noninvertible maps. In order to assist the discussion of the two examples in sections 4.3 and 4.4, we briefly introduce here some necessary concepts from the theory of noninvertible maps; see, for example, [1] for more details.

For simplicity, we consider only planar maps, so we suppose that f is a smooth noninvertible map on $\mathbb{R}^2$. Typically, the phase space is divided into regions where points have different numbers of pre-images. The curves separating these regions are defined as follows. Adopting the notation in [24], we define the so-called *curve of merging pre-images* (also denoted LC_{-1}),

$$J_0 = \{x \mid Df\,(x) \, \text{is singular}\},$$

where the determinant of the Jacobian vanishes. The image of this curve, $J_1 = f(J_0)$, divides the phase space into regions with a constant number of pre-images. (This curve, J_1, is also called *ligne critique* (LC) [7, 24].)

Geometrically, we are dealing with the following problem in singularity theory: How is the plane mapped over itself by the smooth map f? The curve J_0 marks where the phase space folds under iteration of f. The fold is mapped to J_1, which marks a change in the number of pre-images. As one crosses J_1 from one region to the next, one generically crosses a fold, and the number of pre-images increases or decreases by two; see, for example, [2]. The simplest case is when J_1 is a smooth curve that divides the plane into exactly two regions, which is the situation we consider in section 4.3. In this situation the map f is often referred to as a $(Z_0 - Z_2)$ map. In this notation Z_i represents the region with i first-rank pre-images [21].

However, J_1 may have self-intersections, in which case more regions with different numbers of pre-images exist. Furthermore, there may be singularities of higher codimension on the curve J_1 corresponding to a kernel of $Df\,(x)$ of dimension larger than one. We encounter this type of more complicated map in section 4.4 that, locally near the primary manifold, features two folds, with a separate sheet of inverses underneath; we classify this map as type $(Z_{1+1}–Z_{3+1}–Z_{1+1})$. The saddle for this map has a negative stable eigenvalue, so that we need to work with the second iterate. The structure for the number of pre-images in the phase space of the second iterate is even more complicated because the Jacobian $Df^2(x)$ is singular when either $Df(x)$ or $Df(f(x))$ is singular. Hence, for the second iterate f^2 we have

$$J_0(f^2) = \{x \mid Df\,(x) \;\text{ or }\; Df\,(f(x)) \, \text{is singular}\} := J_{-1} \cup J_0,$$

where J_{-1} is the set of pre-images of J_0, that is, $J_0 = f(J_{-1})$. Note that J_{-1} may consist of several curves. Similarly,

$$J_1(f^2) = f^2\left(J_{-1} \cup J_0\right) := J_1 \cup J_2,$$

where $J_2 = f^2\left(J_0\right)$.

3. The algorithm.

The SC algorithm grows a one-dimensional branch of the stable set in steps by adding new points according to the local curvature properties of the branch. The difference between the SC algorithm and the algorithm from [17] lies in how the next point is found. Instead of using the inverse f^{-1} for finding the next point, the SC algorithm finds a new point close to the last computed point that maps under f to a piece of the stable set that was already computed. Before we explain this in detail, we recall for completeness the basic idea of growing a branch step by step.

3.1. Growing a branch of the stable set.

The algorithm produces a piecewise-linear approximation of a branch of the stable manifold, or the primary manifold, of $W^s(x_0)$ by computing an ordered list of points $M = \{p_0, p_1, \dots, p_N\}$ at varying distances from each other. The first point p_0 is the saddle x_0. We use a linear approximation for $W^s_{loc}(x_0)$, so the next point p_1 is taken a small distance $\delta > 0$ from p_0 along $E^s(x_0)$. The distance between consecutive points is adjusted according to the curvature of the branch, as explained below.

Suppose that the branch under consideration has been grown up to p_k such that $M = \{p_0, p_1, \dots, p_k\}$. We wish to find the next point p_{k+1} at a distance Δ_k from p_k. The stable set is forward invariant, so new points, and in particular p_{k+1}, must map back onto the piece of the branch that has been computed so far. Indeed, if the map is a diffeomorphism and f^{-1} is known, then we can search the previously computed part of the branch to find a point that maps at distance Δ_k from p_k under f^{-1}. This is the method used in [17] to compute the stable manifold. We wish to avoid the use of f^{-1} and instead use a different method, which is explained in detail in section 3.2.

We use the strategy of [17] to ensure an acceptable resolution of the curve according to prespecified accuracy parameters. In particular, we monitor α_k, the angle between p_{k-1}, p_k, and p_{k+1}, and the product $\Delta_k \alpha_k$. We approximate α_k by

$$\alpha_k = 2\sin^{-1}\left(\frac{\|\bar{p} - p_{k-1}\|}{2\|p_k - p_{k-1}\|}\right) \approx \frac{\|\bar{p} - p_{k-1}\|}{\|p_k - p_{k-1}\|},$$

where

$$\bar{p} = p_k + \frac{\|p_k - p_{k-1}\|}{\|p_k - p_{k+1}\|}\left(p_k - p_{k+1}\right)$$

is the point on the line through p_k and p_{k+1} that lies at the same distance from p_k as p_{k-1}. We check the conditions

$$\alpha_k < \alpha_{\max}, \tag{3.1}$$

$$\Delta_k \alpha_k < (\Delta\alpha)_{\max}. \tag{3.2}$$

Condition (3.1) ensures that enough points are computed along the branch, and condition (3.2) controls the local interpolation error. If one of these criteria is not satisfied, then we

```
Pseudocode

Grow_Manifold (Fixed point: $p_0$, first point along $E^s(p_0)$: $p_1$, target arclength: $A$)
    $M = \{p_0, p_1\}$;
    $p_{\text{left}} = p_0$;
    $p_{\text{right}} = p_1$;
    $arclength = \|p_1 - p_0\|$;
    while ($arclength < A$)
        $p_k$ = last point in $M$;
        $p_{k-1}$ = next to last point in $M$;
        $(p_{\text{candidate}}, p_{i-1}, p_i)$ = Search_Circle($M$, $p_{\text{left}}$, $p_{\text{right}}$);
        /* $[p_{i-1}, p_i]$ is the interval in which $f(p_{\text{candidate}})$ lies. */
        $\alpha_k = \angle(p_{k-1}, p_k, p_{\text{candidate}})$
        /* $\alpha_k$ is the angle between $p_{k-1}$, $p_k$ and $p_{k+1}$. */
        if ( ($\alpha_k < \alpha_{\max}$ and $\Delta_k\alpha_k < (\Delta\alpha)_{\max}$) or $\Delta_k < \Delta_{\min}$)
            /* Accept point. */
            $p_{k+1} = p_{\text{candidate}}$;
            append $p_{k+1}$ to $M$;
            $arclength = arclength + \Delta_k$;
            $p_{\text{left}} = p_{i-1}$;
            $p_{\text{right}} = p_i$;
            if ($\alpha_k < \alpha_{\min}$ and $(\Delta_k\alpha_k) < (\Delta\alpha)_{\min}$)
                /* Increase $\Delta_k$ for the next step. */
                $\Delta_k = 2\Delta_k$;
            end if
        else
            /* Accuracy conditions not satisfied. Reject point and decrease $\Delta_k$. */
            $\Delta_k = \Delta_k/2$;
        end if
    end while
    return $M$;
end.
```

Figure 1. *A pseudocode representation of the* Grow_Manifold *algorithm, which calls* Search Circle *in Figure 3 to find the next point on the branch.*

replace Δ_k with $\frac{1}{2}\Delta_k$ and repeat the procedure to find a new candidate for p_{k+1}. Otherwise, we accept p_{k+1} and set $\Delta_{k+1} = \Delta_k$, unless both $\alpha_k < \alpha_{\min}$ and $\Delta_k\alpha_k < (\Delta\alpha)_{\min}$ hold for a user-specified choice of parameters $\alpha_{\min}$ and $(\Delta\alpha)_{\min}$. If this is the case, then the manifold is relatively straight and we try $\Delta_{k+1} = 2\Delta_k$. This ensures that the number of points used to approximate the branch is in some sense optimized for the required accuracy constraints. Note that, at sharp folds, it may be necessary to accept $\alpha_k > \alpha_{\max}$ due to Δ_k becoming very small. In this case we accept the point if $\Delta_k < \Delta_{\min}$, where $\Delta_{\min}$ is also a predefined parameter, and we note the occurrence; see [17] for more details.

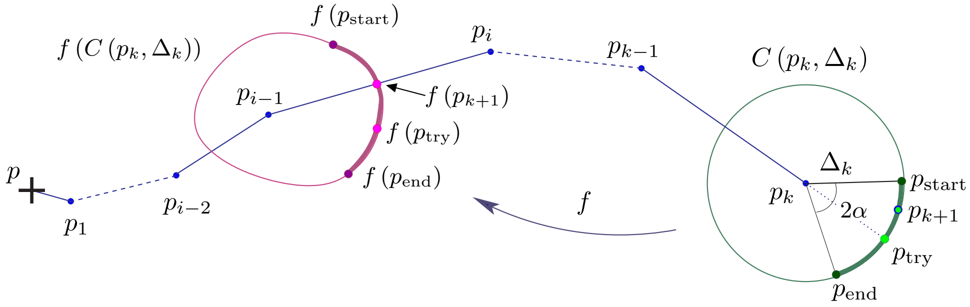


Figure 2. *Graphical illustration of the SC algorithm. A new point p_{k+1} is found on the circle $C(p_k,\Delta_k)$ centered at p_k with radius Δ_k, such that $f(p_{k+1})$ lies on a previously computed part of $W^s(x_0)$.*

A pseudocode representation of how the branch is grown is given in Figure 1.

3.2. The SC method. We now explain how the next point along $W^s(x_0)$ is found when f^{-1} is unknown. The SC algorithm finds the point p_{k+1} at distance Δ_k from p_k as a point that maps back onto a piece of branch that was already computed. The method is illustrated in Figure 2 and described in pseudocode in Figure 3. We know that p_{k+1} must lie somewhere on a circle of radius Δ_k centered at p_k. The image of this circle must then intersect a previously computed part of the branch. Note that the circle centered at p_k itself necessarily intersects the branch that has been computed so far. Hence, its image intersects the branch at the image of this intersection point. If Δ_k is small enough, there will be a unique second intersection, which is the image of the point p_{k+1} that we wish to find. As depicted in Figure 2, the point p_{k+1} must lie somewhere on the green circle $C(p_k,\Delta_k)$. The image $f(C(p_k,\Delta_k))$ of this circle is the closed loop shown in magenta. If Δ_k is small enough, then $f(C(p_k,\Delta_k))$ will only intersect the branch twice; we are not interested in one of the intersections because it is the image of a point on the previously computed branch. The other point intersects the segment, say $[p_{i-1},p_i]$, of the approximated branch, and it is the image of p_{k+1}, as indicated by the arrow. The point that maps to this intersection is our candidate for p_{k+1}.

The accuracy constraint (3.1) puts an upper bound on the allowable angle between consecutive segments. We can ensure immediately that α_k does not exceed $\alpha_{\max}$ by only searching the part of the circle that satisfies this criterion. We denote this search region by $C_\alpha(p_k,\Delta_k)$, which is the arc between p_{start} and p_{end}; see the segment of the green circle in Figure 2 enclosed by the black lines. The image of this segment is the thick magenta curve $f(C_\alpha(p_k,\Delta_k))$. Note that this automatically ensures that we do not accidentally search for a pre-image of the other intersection point (which we are not interested in).

In order to find p_{k+1} we first need to find the segment $[p_{i-1},p_i]$ of the calculated branch that contains the intersection point with $f(C_\alpha(p_k,\Delta_k))$. We start this search with the segment that was used in the previous step to find a candidate for p_{k+1}, which we denote $[p_{\text{left}},p_{\text{right}}]$.

An important step in this search is performed in the routine FIND_POINT_ON_LINE; see Figure 3. We use bisection on $C_\alpha(p_k,\Delta_k)$ to find a point p_{try} on the circle that maps within a normal distance ε_B to the line $\{(1-\tau)\,p_{\text{left}}+\tau p_{\text{right}} \mid \tau\in\mathbb{R}\}$; that is, we allow for the

```
Search_Circle ($M$, $p_{\mathrm{left}}$, $p_{\mathrm{right}}$)
    do
        ($p_{\mathrm{candidate}}, \tau$) = Find_Point_On_Line ($p_{\mathrm{left}}$, $p_{\mathrm{right}}$);
        /* If $\tau < 0$ or $\tau > 1$, point is on line, but not on segment. */
        if ($\tau < 0$) /* move backward. */
            $(p_{\mathrm{left}}, p_{\mathrm{right}}) = (p_{\mathrm{left}-1}, p_{\mathrm{left}})$;
        else if ($\tau > 1$) /* move forward. */
            $(p_{\mathrm{left}}, p_{\mathrm{right}}) = (p_{\mathrm{right}}, p_{\mathrm{right}+1})$;
        end if
    while ($\tau \notin [0,1]$);
    return ($p_{\mathrm{candidate}}, p_{\mathrm{left}}, p_{\mathrm{right}}$);
end.

Find_Point_On_Line ($p_{\mathrm{left}}$, $p_{\mathrm{right}}$)
    $L(\tau)$ = line $[p_{\mathrm{left}}, p_{\mathrm{right}}] = \{(1-\tau)p_{\mathrm{left}} + \tau p_{\mathrm{right}} \mid \tau \in \mathbb{R}\}$;
    $\theta_{\mathrm{start}} = -\alpha_{\mathrm{max}}$;
    $\theta_{\mathrm{end}} = \alpha_{\mathrm{max}}$;
    $p_{\mathrm{start}}$ = point on circle at angle $\theta_{\mathrm{start}}$;
    $p_{\mathrm{end}}$ = point on circle at angle $\theta_{\mathrm{end}}$;
    $\bar{p}$ = point on circle at angle 0;
    $\vec{V}_{\mathrm{start}} = f\,(p_{\mathrm{start}}) - p_{\mathrm{right}}$;
    $\vec{V}_{\mathrm{end}} = f\,(p_{\mathrm{end}}) - p_{\mathrm{right}}$;
    $\vec{W}$ = normal vector to $p_{\mathrm{left}} - p_{\mathrm{right}}$;
    if ( $\langle \vec{V}_{\mathrm{start}}, \vec{W} \rangle * \langle \vec{V}_{\mathrm{end}}, \vec{W} \rangle > 0$ )
        /* $f(p_{\mathrm{start}})$ and $f(p_{\mathrm{end}})$ do not lie on opposite sides of $L$. */
        increase $C_\alpha\,(p_k, \Delta_k)$ search region.
    end if
    do
        /* Bisection to find point on $L(\tau)$. */
        $\theta_{\mathrm{try}} = (\theta_{\mathrm{end}} + \theta_{\mathrm{start}})\,/2$;
        $p_{\mathrm{try}}$ = point on circle $C(P_k, \Delta_k)$ at angle $\theta_{\mathrm{try}}$ from $\bar{p} - p_k$.
        $\vec{V} = f\,(p_{\mathrm{try}}) - p_{\mathrm{right}}$;
        if ( $\langle \vec{V}_{\mathrm{end}}, \vec{W} \rangle * \langle \vec{V}, \vec{W} \rangle > 0$ )
            /* $f(p_{\mathrm{try}})$ is on same side as $f(p_{end})$. */
            $\theta_{\mathrm{end}} = \theta_{\mathrm{try}}$;
        else
            /* $f(p_{\mathrm{try}})$ is on same side as $f(p_{start})$. */
            $\theta_{\mathrm{start}} = \theta_{\mathrm{try}}$;
        end if
    while ( $|\langle \vec{V}, \vec{W} \rangle| < \varepsilon_B$)
    /* Normal distance between $L$ and $f\,(p_{\mathrm{try}}) < \varepsilon_B$.
        Accept as candidate.*/
    return ($p_{\mathrm{try}}, \tau$).
end.
```

Figure 3. *The* Search Circle *routine, given here in pseudocode, finds a new point p_{k+1} on a branch of the stable set at distance Δ_k from p_k, the last point computed so far.*

possibility that $f(p_{\text{try}})$ is only ε_B-close to the closest point $(1-\tau)p_{\text{left}} + \tau p_{\text{right}}$ on the line. Here ε_B is a prespecified bisection tolerance; see section 3.6 for a discussion on the bisection. To start the bisection we require that p_{start} and p_{end} lie on opposite sides of the line, which we check by assessing the sign of the projection of these points onto the normal of the line. The next point p_{try} is then found by bisection, and in the next bisection step it replaces the point that lies on the same side of the line. The bisection stops when the normal distance from $f(p_{\text{try}})$ to the segment $[p_{\text{left}}, p_{\text{right}}]$ is less than ε_B, and it also returns the value of τ corresponding to the projection of $f(p_{\text{try}})$ onto the line.

We now need to check that p_{try} actually maps inside the segment $[p_{\text{left}}, p_{\text{right}}]$; that is, we need to check that $0 \leq \tau \leq 1$. If this is indeed the case, then p_{try} is a candidate for the next point p_{k+1}. If $\tau > 1$, we repeat the search with the following segment $[p_{\text{right}}, p_{\text{right}+1}]$. If the map has multiple pre-images, then, as we search for new points to add to the branch, their images may start to move backward along the branch. In this case, $\tau < 0$, and we need to search for $f(p_{k+1})$ on the previous segment $[p_{\text{left}-1}, p_{\text{left}}]$; see the examples of noninvertible maps in sections 4.3 and 4.4.

Once we have found a proper candidate for p_{k+1} whose image lies within a normal distance ε_B to $\{(1-\tau)\, p_{\text{left}} + \tau p_{\text{right}} \mid 0 < \tau < 1\}$, we return the candidate $p_{\text{candidate}}$, together with the respective segment, which is the segment $[p_{i-1}, p_i]$ that we set out to find. We then check the accuracy conditions (3.1) and (3.2). If these conditions are satisfied, then the candidate point is added as p_{k+1} to the list of points M, and Δ_k is adjusted if necessary. If the candidate fails the accuracy conditions, Δ_k is halved and the entire procedure is repeated, as was explained in section 3.1.

Note that, occasionally, p_{start} and p_{end} may not be on opposite sides of the segment $[p_{\text{left}}, p_{\text{right}}]$. In this case we increase α_k beyond its allowed maximum until the starting conditions for the bisection procedure are satisfied. Normally, this happens only if there is a sharp fold in the stable set or if we are searching for p_{k+1} using the incorrect segment. In the latter case, the point will be rejected anyway. Otherwise, we accept the point and give a warning message, provided $\Delta_k < \Delta_{\min}$; see [17] for a further discussion of sharp folds.

Figure 4 shows the algorithm in practice, computing the stable manifold for the shear map of section 4.1. Figure 4(a) shows the algorithm searching for p_{k+1}, with the points p_{start} and p_{end} and the successive points p_{try} of the bisection in green. Figure 4(b) shows the images of these search points, getting closer to the segment $[p_{\text{left}}, p_{\text{right}}]$. Bisection is done until $f(p_{\text{try}})$ lies at distance $\varepsilon_B = 10^{-6}$ from $[p_{\text{left}}, p_{\text{right}}]$. As shown in Figure 4, $f(p_{\text{try}})$ lies ε_B-close to the segment $\{(1-\tau)p_{\text{left}} + \tau p_{\text{right}} \mid 0 \leq \tau \leq 1\}$, so p_{try} is a proper candidate for p_{k+1} and $[p_{i-1}, p_i] = [p_{\text{left}}, p_{\text{right}}]$. Furthermore, p_{try} satisfies conditions (3.1) and (3.2), and thus it is accepted and the new segment $[p_k, p_{k+1}]$, with $p_{k+1} = p_{\text{try}}$ is added to the approximation of $W^s(x_0)$.

3.3. Estimating the initial value for Δ_1. When we start our algorithm from the saddle point, the point p_1 is taken at a small distance δ along the stable eigendirection, $E^s(x_0)$. It is not straightforward what value to choose for Δ_1. This distance must be much smaller than restricted by the accuracy criteria, because the primary manifold stretches relatively little close to the fixed point. We cannot just iterate the endpoints, as we do not know the inverse, but we wish to grow the primary manifold such that $f(p_{k+1}) \approx p_k$.

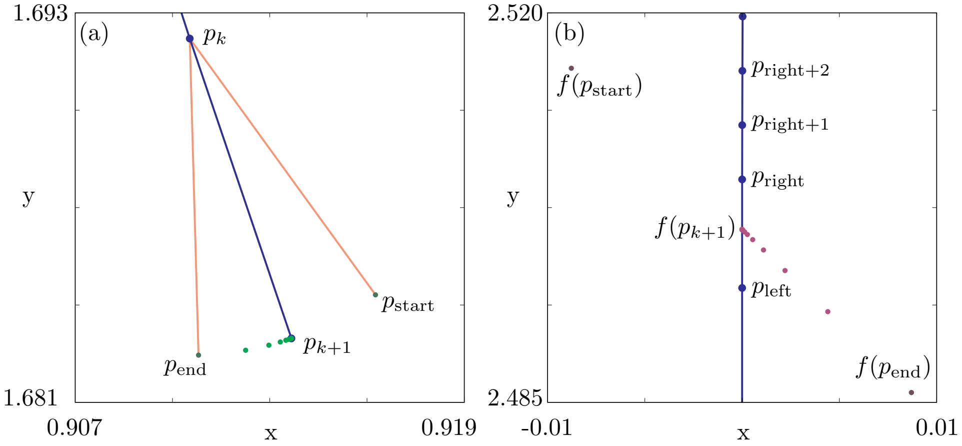


Figure 4. *Computing the stable manifold for the shear map of section 4.1. Searching $C_\alpha(p_k, \Delta_k)$ using bisection to find p_{k+1}* (a) *and the images of the green points on $C_\alpha(p_k, \Delta_k)$ in* (a) *on $f(C_\alpha(p_k, \Delta_k))$* (b)*. Only $f(p_{k+1})$ also lies on $[p_{i-1}, p_i]$. Therefore, only $p_{k+1} \in C_\alpha(p_k, \Delta_k)$ lies on $W^s(x_0)$.*

Hence, our aim is to add a new point p_2 at distance Δ_1 from p_1, such that the point $f(p_2)$ lies on $[x_0, p_1]$ very close to p_1. If we choose Δ_1 too small, then the algorithm computes too many mesh points close to the saddle. However, if we choose Δ_1 too large, then the image of the segment $C_\alpha(p_k, \Delta_k)$ does not intersect $[x_0, p_1]$ at all. A good estimate for Δ_1 is found as follows. The stable eigenvalue λ_s is the linear contraction rate along the stable eigenvector. Near the saddle, we may use λ_s as an estimate of the nonlinear contraction, and thus the initial estimate is

$$\lambda_s(\delta_1 + \Delta_1) = \delta_1 \Leftrightarrow \Delta_1 = \delta_1 \left(\frac{1}{\lambda_s} - 1 \right),$$

where $\delta_1 = \delta$ is the distance from x_0 to p_1. Since Δ_1 may still be too large, we search over the interval $\Delta_1 \pm \frac{1}{2}\Delta_1$ on a line through x_0 and p_1 to find an acceptable Δ_1. We use bisection on points in this interval to find points that map close to p_1, assessing on which side the points lie of a line perpendicular to the line through x_0 and p_1. We do not have to be exactly on this line, but we have to ensure that we do not choose Δ_1 too large. This method may need to be applied several times to find acceptable Δ_k's until the accuracy conditions are met. Then the normal SC algorithm is started.

3.4. Maps with multiple inverses. For a map with multiple pre-images, the SC algorithm works in exactly the same way, except that it is now possible for the images of the newly added points p_{k+1} to traverse back along the manifold toward the saddle. This happens when the manifold crosses the curve of merging pre-images J_0; see section 2.1 for details. This phenomenon is automatically detected by our algorithm and does not cause any problems. In section 4.3 we discuss this in more detail.

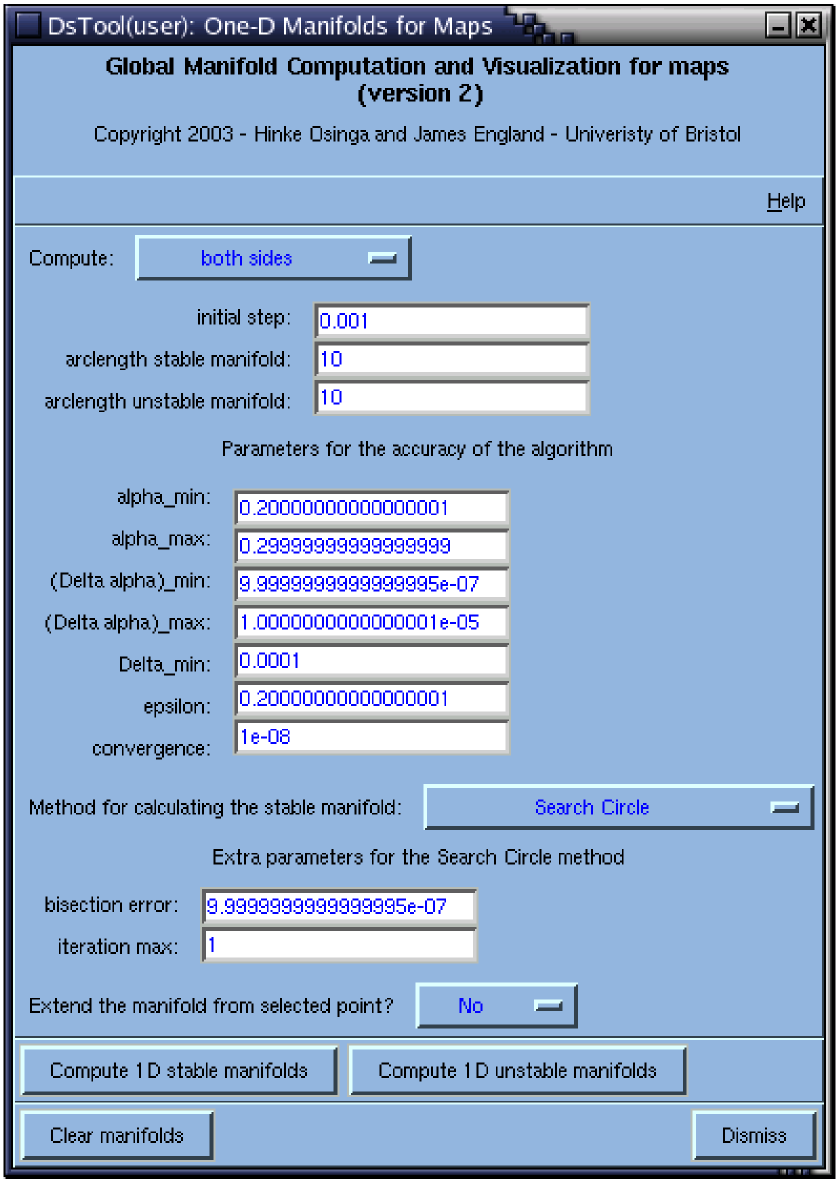
If part of the stable set lies in a region with at least three pre-images, then it is possible that the image of one branch of the primary manifold covers a part of the other branch; an example of this phenomenon is investigated in section 4.4. (This should not be confused with the case in which the eigenvalues of the saddle are negative, where one branch of the manifold is mapped to the other.) In this case the images of points that are added to the branch traverse back along the computed branch until the fixed point is reached again. That is, the new point that is added to the computed branch is another pre-image of the saddle point, which lies by definition on $W^s(x_0)$. At this moment, further new points map to the other side of the saddle, that is, they lie on the branch that has not yet been computed. Therefore, we must stop and compute part of the other branch first before the present branch can be grown further. It may be necessary to stop several times and swap between the growth of the two branches to grow one or both up to the required arclength.

In noninvertible maps there may be disjoint parts of the stable set. The SC algorithm can also be used to find these. This is done by selecting, for example, pre-images of the saddle point that do not lie on the branch computed so far. Because the saddle point is not on the curve J_0, the Jacobian of the saddle point is nonsingular and, for example, Newton's method can be used to find its pre-images. The next point on the respective branch is then found by allowing a search of the whole circle for points that map back onto the primary manifold (which contains the saddle point). At this initial step we use $\Delta_k = \delta$. Note that, when computing a disjoint branch, the SC method can start as soon as *one* point on the new branch has been found, meaning that no fundamental domain needs to be constructed; see the example of the noninvertible map in section 4.4.

3.5. DsTool implementation. The SC algorithm has been fully implemented in the Tcl/Tk version of the DsTool environment [3]. The code can be downloaded as an extension module to DsTool and is available with this paper; see the link in [26]. The installation procedure is straightforward and instructions are included with the code.

Note that our module can be viewed as a replacement of the module presented in [19], which implements the algorithm in [17]. The implementation of the SC algorithm has been combined with that algorithm so that the unstable manifold can be computed with the method of [17] and the stable set (initially, the primary manifold) using either method, depending on whether the inverse is known or not. The window that appears in DsTool when using the module allows the user to choose which method to use to calculate the stable manifold or set. In fact, it is possible to directly compare the two algorithms in cases where f^{-1} is known or approximated by Newton's method.

A screenshot of the module window in DsTool is shown in Figure 5. The window is essentially the same as that presented in [19] but with some additional features incorporating the SC algorithm. As before, the user can decide which branch to compute: *positive side*, *negative side*, or *both sides*, using the selection menu at the top. Below that, one can specify the initial step size δ, and the required arclength of both the stable and unstable manifold computations. (Note that it is now possible to compute only, say, the stable manifold without setting the arclength for the unstable manifold to zero.) Next, the accuracy parameters $\alpha_{\min}$, $\alpha_{\max}$, $(\Delta\alpha)_{\min}$, $(\Delta\alpha)_{\max}$, and $\Delta_{\min}$ that are used by both algorithms are specified; see section 3.1. The parameter ε is the uncertainty factor from [17]. The parameter *convergence* controls the detection of convergence to a fixed point, which happens when the branch has finite arclength.


DsTool(user): One-D Manifolds for Maps
Help
Compute 1D stable manifolds
Compute 1D unstable manifolds
Clear manifolds
Dismiss

Global Manifold Computation and Visualization for maps (version 2)

Copyright 2003 - Hinke Osinga and James England - Univeristy of Bristol

Compute: both sides

initial step: 0.001
arclength stable manifold: 10
arclength unstable manifold: 10

Parameters for the accuracy of the algorithm

alpha_min: 0.20000000000000001
alpha_max: 0.29999999999999999
(Delta alpha)_min: 9.9999999999999995e-07
(Delta alpha)_max: 1.0000000000000001e-05
Delta_min: 0.0001
epsilon: 0.20000000000000001
convergence: 1e-08

Method for calculating the stable manifold: Search Circle

Extra parameters for the Search Circle method

bisection error: 9.9999999999999995e-07
iteration max: 1

Extend the manifold from selected point? No

Figure 5. *The implementation of the SC algorithm as it appears in the DsTool environment.*

The user is now able to select the method from the drop-down menu that should be used to compute the stable manifold or set. The options are the *Explicit/Approximate inverse* method, the *Monte Carlo* method, and the Search Circle method. If the *Explicit/Approximate inverse* or *Monte Carlo* method is selected, then the algorithm from [17] will be used to compute the stable manifold. The Monte Carlo method uses random seeds for Newton's method to find the inverse, whether an inverse has been defined or not. If no explicit or approximate inverse is defined, then the *Explicit/Approximate inverse* option will not appear in the drop-down menu.

If the Search Circle method is selected, then two extra parameters appear. The bisection error ε_B is set in *bisection error*. The parameter *iteration max* is only used when the map has multiple inverses; see section 4.4 for more details. The final drop-down menu allows the user to extend the stable set from a selected point. If this option is selected, then the algorithm will try to extend disjoint parts of the stable set from points saved in the DsTool *Selected* window. This is important for cases where the map has multiple inverses and disjoint pieces of the stable set exist. The user must first find a point on a disjoint piece. This can be done, for example, by using Newton's method in combination with Monte Carlo seeds for finding pre-images of the fixed points that do not lie on the pieces computed so far.

3.6. Comment on accuracy. A complete discussion of the accuracy of the basic algorithm is given in [13] and [17]. Our method introduces an additional error, which does not appear in [17], due to the bisection error ε_B. This is a parameter of the computation, and we can make it as small as we like. It is not practical to set this parameter to zero, but by setting it to the same order as the interpolation error, the overall error is of the order given in [17].

4. Examples. In this section we use the SC algorithm to compute approximations of the stable manifolds or sets of five planar maps. The same accuracy parameters are used throughout, unless otherwise stated. The initial step is $\delta = 10^{-3}$ and the other accuracy parameters are $\alpha_{\min} = 0.2$, $\alpha_{\max} = 0.3$, $(\Delta\alpha)_{\min} = 10^{-6}$, $(\Delta\alpha)_{\max} = 10^{-5}$, and $\Delta_{\min} = 10^{-4}$. We chose a bisection error $\varepsilon_B = 10^{-6}$ for all calculations.

We begin with the shear map in section 4.1. The shear map is an ideal test case for all algorithms computing stable and unstable manifolds. Namely, the local stable and unstable manifolds are parts of the coordinate axes of the system, and there is a homoclinic tangency for a specific parameter value. Therefore, we know that the algorithm is accurate if the folds of the manifolds are tangent to the axes in a neighborhood of the origin.

Our next example, in section 4.2, is a modified form of the Ikeda map. It was introduced in [16] as an example of a diffeomorphism that does not have an inverse in closed analytic form. We use this map for comparison with the algorithm in [16], which uses Newton's method to approximate the inverse.

Two different maps with multiple inverses are considered in sections 4.3 and 4.4. In section 4.3 the stable set is such that the primary manifold forms a closed loop; there may also be a finite number of disjoint pieces of the stable set. In section 4.4 the stable set consists of infinitely many disjoint curves and, starting from the saddle point, our algorithm computes the primary manifold without difficulty; the disjoint parts can be computed as well by starting from respective pre-images of the fixed point.

Finally, we discuss a model of a highly interrupted cutting process in section 4.5. Most real-life systems are either noninvertible or difficult to invert due to their complexity. For this map, which is the real-life mechanical system describing a cutting tool [4, 5, 33], the Jacobian of the system is discontinuous along a straight line in phase space. (The map can be split into two diffeomorphisms, each having a unique inverse, and the manifolds are continuous across the smoothness boundary.) We compute the stable and unstable piecewise-smooth manifolds even though they cross the smoothness boundary. To our knowledge, global manifolds of piecewise-smooth systems have not been computed before.

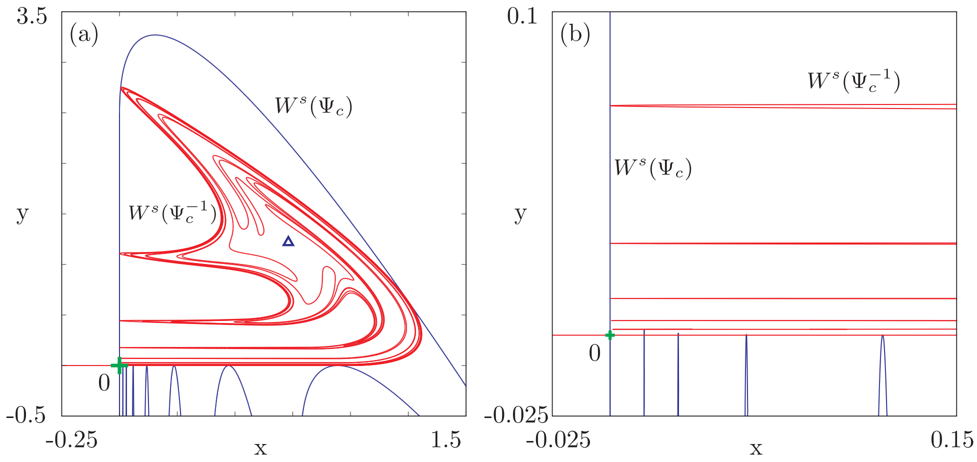


Figure 6. *The shear map computed at first tangency* ($c = 0.811580$) *with the stable manifold calculated using the SC method. The unstable manifold was computed as the stable manifold of the inverse of the shear map, also using the SC method. Panel* (b) *is an enlargement near the fixed point. (The main branch of* $W^s(0)$ *was computed to arclength* 5000 *and* $W^u(0)$ *to arclength* 100.*)*

4.1. Shear map. The shear map was first introduced in [17] and is based on an abstract construction given in [28]. The map is such that the origin is always a saddle point and the local stable and unstable manifolds are the coordinate axes of the system.

The shear map is a composition of simple linear saddle-type behavior with a nonlinear shear that is effective only away from a neighborhood of the origin. The linear map is defined as

$$\phi \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda^u x \\ \lambda^s y \end{pmatrix},$$

where the stable and unstable eigenvalues $0 < \lambda^s \leq (\lambda^u)^{-1} < 1$ are fixed parameters. The nonlinear shear ψ_c depends on the parameter c and is defined as

$$\psi_c \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - c\, s(x+y) \\ y + c\, s(x+y) \end{pmatrix},$$

where

$$s(z) = \begin{cases} 0 & \text{for } z \leq 1, \\ (z-1)^2 & \text{for } z > 1. \end{cases}$$

Hence ψ_c produces a quadratic shear along the diagonals $x + y = \text{const} > 1$. The shear map is defined as the composition

$$\Psi_c = \psi_c \circ \phi.$$

By construction, $W^s(\mathbf{0})$ contains the half-line $\{x \in (-\infty, 1], y = 0\}$ and $W^u(\mathbf{0})$ contains $\{y \in (-\infty, 1/\lambda^s], x = 0\}$. There is a value $c = c^* \in [0, \lambda^u]$ for which there is an initial homoclinic tangency. The homoclinic tangency for $\lambda^u = 0.4$ and $\lambda = 2.0$ was found to be $c^* \approx 0.811580$, which is precise up to five digits; see [17]. Figure 6 shows the stable and

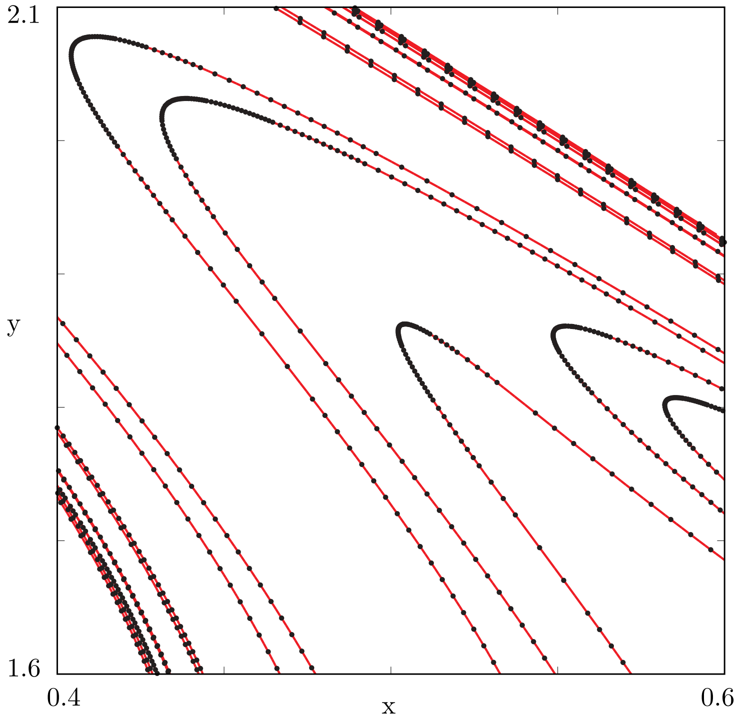


Figure 7. *A part of $W^u(\mathbf{0})$ (red curve) overlaid with black points computed with the SC algorithm as the stable manifold of the inverse map.*

unstable manifolds of the shear map, both computed using the SC algorithm at this homoclinic tangency. Here Ψ_c was used to calculate $W^s(\mathbf{0})$, shown in blue, and Ψ_c^{-1} was used to calculate $W^u(\mathbf{0})$, shown in red. That is, we calculated $W^u(\mathbf{0})$ as the stable manifold of Ψ_c^{-1} (note that one would not normally use the inverse to calculate an unstable manifold, but we did this for testing purposes because $W^u(\mathbf{0})$ is bounded, while $W^s(\mathbf{0})$ is not).

Figure 6(b) shows the results of our computation, where we zoomed in near the fixed point, a test that was also performed in [17]. There are no visible differences in the two computations. We remark that the stable manifold makes longer and longer excursions into the region of negative y before returning to the next tangency. However, due to the way the mesh is adapted, only very few mesh points are computed along each such excursion; see the discussion in [17] and also [35]. Note that the limit of numerical accuracy was reached near the fixed point: the last fold of $W^s(\mathbf{0})$ crosses $W^u(\mathbf{0})$, but the last fold of $W^u(\mathbf{0})$ just misses $W^s(\mathbf{0})$.

Figure 7 shows in red a close-up piece of $W^u(\mathbf{0})$ calculated using Ψ_c and the algorithm from [17]. The black points are on the stable manifold of Ψ_c^{-1} computed with the SC algorithm. Here we used the larger values $(\Delta\alpha)_{\min} = 10^{-5}$ and $(\Delta\alpha)_{\max} = 10^{-4}$ to demonstrate how the distribution of points is adapted to the curvature. Despite the lower accuracy, the black points computed with the SC method lie exactly on the red manifold.

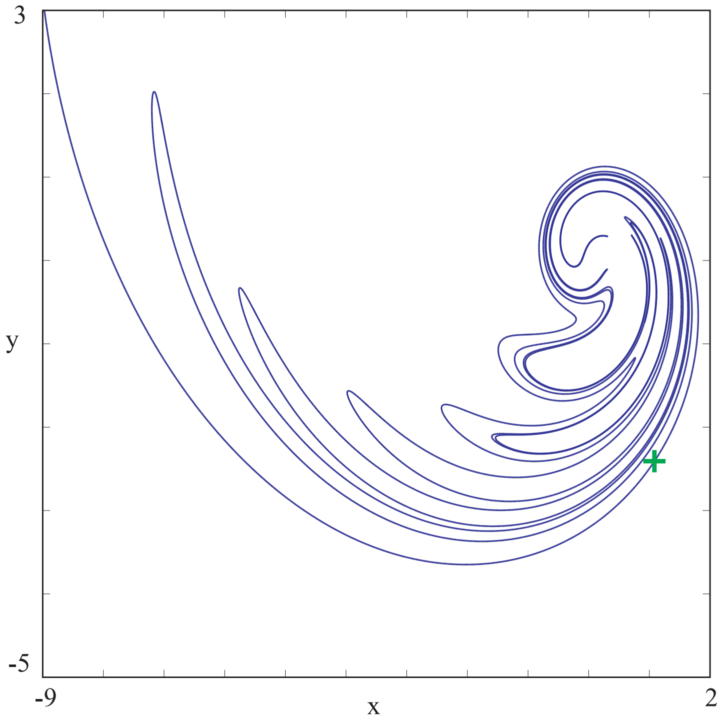


Figure 8. *The stable manifold of the modified Ikeda map up to arclength* 150 *calculated using the SC method. ($A = 1, b = 0.9, e = 1, \phi = 0.4, q = 6$.)*

4.2. Modified Ikeda map. The Ikeda map describes the dynamics of an optical ring laser cavity [12, 15]. A modification of the Ikeda map is used in [16] as an example of a system that does not have a closed analytic form of the inverse map. It is defined by

$$g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} A + bx\cos m - ey\sin m \\ by\cos m + ex\sin m \end{pmatrix}, \tag{4.1}$$

where

$$m = \phi - \frac{q}{1 + x^2 + y^2},$$

and A, b, e, ϕ, and q are parameters of the system. The standard Ikeda map has $b = e$, and it is possible to find an explicit inverse. We consider $b = 0.9$, $e = 1$ and use $A = 1$, $\phi = 0.4$, and $q = 6$ as was done in [16]. Figure 8 shows the stable manifold of the saddle $(1.08332, -2.40796)$, indicated by a green cross, computed up to arclength 150. The two branches of the stable manifold are shown in blue; one of the branches immediately goes to infinity; the other branch stretches and folds as it forms a homoclinic tangle creating a chaotic saddle, while at the same time also going to infinity. This branch is computed to approximately the same arclength as shown in [16, Figure 3], where 17 iterates of a fundamental domain were computed and the inverse was approximated by Newton's method. (The exact fundamental domain is not specified in [16].) There appear to be no differences between the approximations of the stable manifold in Figure 8 and [16, Figure 3].

4.3. Modified Gumowski–Mira map. In [10, 11] Gumowski and Mira investigated a noninvertible map for which the two sides of the primary manifold of the stable set of a saddle point join to form a smooth closed loop; see also [8] for a different example. The area enclosed by the primary manifold is the basin of attraction of a sink. Here we study a modified form of this map to demonstrate the simplest case of this phenomenon, namely, of a saddle point with positive eigenvalues, where the orientation of the map is preserved. The map is defined as

$$Q\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ ax + bx^2 + y^2 \end{pmatrix}. \tag{4.2}$$

For $a = 0.8$ and $b = 1$ we obtain the system studied in [10, 11]. However, we choose a and b such that the map Q has a saddle fixed point p with two positive eigenvalues.

The determinant of the derivative vanishes along the vertical line $J_0 := \{x = -a/2b\}$. The curve $J_1 = Q(J_0)$ is then a parabola and divides the plane into two regions, one with two pre-images and the other with no pre-images. This is the simplest case of a noninvertible map of so-called $(Z_0 - Z_2)$ type; see also section 2.1. The saddle p always has at least one pre-image, namely itself, so that it cannot lie in the Z_0 region. Hence, p lies in the Z_2 region. Therefore, p has a second pre-image, not equal to p, which lies in the Z_0 region. For this particular example, the second pre-image of p, denoted q, has the same y-coordinate as p.

The computation of the stable set $W^s(p)$ for this map is difficult because it crosses the curve J_0, where the Jacobian is singular. Algorithms that use Newton's method, such as in [16], will fail to compute $W^s(p)$ in its entirety. At best, only the part of $W^s(p)$ up to J_0 can be computed. Gumowski and Mira [10, 11] computed $W^s(p)$ for $a = 0.8$ and $b = 1$ by using fundamental domain iteration, which requires detailed knowledge of the two inverses.

The SC algorithm is more general and does not need any specific information about the system. Two examples of a stable set are shown in Figure 9. In Figure 9(a) we chose $a = -0.8$ and $b = 0.2$ and in Figure 9(b) we used $a = -0.8$ and $b = 0.1$. Notice that the stable set $W^s(p)$ is symmetric about J_0 and that the primary manifold forms a closed loop: its two branches meet at the pre-image q of the saddle point p.

Let us first focus our attention on Figure 9(a). Most of the phase space in this figure belongs to the Z_0 region, where no pre-images exist. Only the points above the parabola J_1 have two pre-images and belong to the Z_2 region. Even though it is not clear from the figure, p does lie inside this region. Both J_0 and J_1 intersect $W^s(p)$ exactly twice. The (symmetric) part of $W^s(p)$ to the right of J_0 is mapped by Q onto the part of $W^s(p)$ in between the two intersections with J_1. The part to the left of J_0 is contracted to this same segment, which accounts for the two pre-images of $W^s(p)$ in this region. Figure 9(c) gives a schematic representation of how the two branches of $W^s(p)$ in Figure 9(a) are mapped by Q. (Here the primary manifold is "cut" at q and "stretched out" to a straight line with the saddle p as the cross in the center.) Also marked are the intersections with the J_0 and J_1 curves. Consecutive sections are indicated by alternating colors, green for the left branch and blue for the right branch. The corresponding images of these sections are shown above and below the primary manifold.

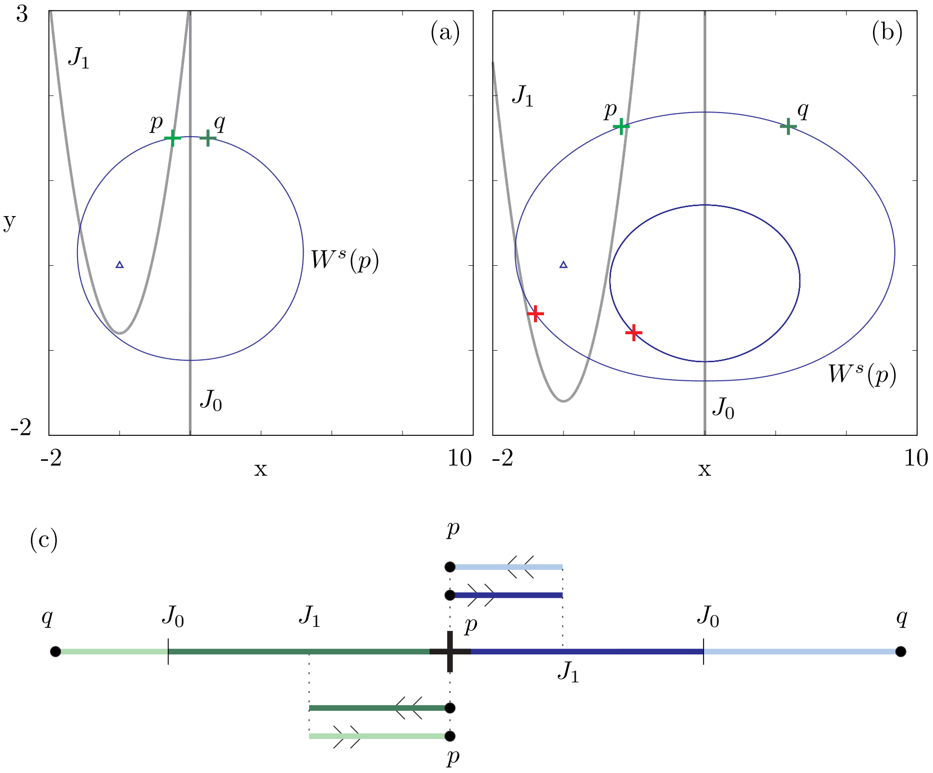


Figure 9. *Panel* (a) *shows the primary manifold of the map defined in* (4.2) *for* $a = -0.8$ *and* $b = 0.2$*, computed up to arclength* 20*. Panel* (b) *shows the stable set of the map defined in* (4.2) *for* $a = -0.8$ *and* $b = 0.1$*. The red crosses indicate a point on the primary manifold and its pre-image, from which the disjoint piece of the stable set was grown. Panel* (c) *is a schematic illustration of how the primary manifold of panel* (a) *is grown. Segments of a particular color map to similarly colored segments.*

Figure 9(c) also illustrates how the SC algorithm computes the primary manifold. Let us consider the right branch of the primary manifold in Figure 9(c). The SC algorithm grows the branch starting with p and a point in $E^s(p)$ at distance $\delta = 10^{-3}$ from p. The part of the manifold from p up to the intersection with J_0 is mapped one-to-one to the segment from p up to the intersection with J_1, which is indicated by the similarly colored piece shown immediately above the primary manifold. The intersection with J_0 may be called a *point of alternance* [11], since here the direction of the images switches. This is detected by our algorithm by a change in direction of the search for the image of p_{k+1}. Namely, the segment from the intersection with J_0 up to q is mapped to the piece of the primary manifold that traverses back from J_1 to the saddle point p. Further points on the branch past q would then map onto the left branch of the primary manifold that has not yet been computed, and so the algorithm stops. The left branch is grown in the same way. Note that the two branches connect smoothly at q because the image of $W^s(p)$ near q under the smooth map Q is the part of $W^s(p)$ near p, which is indeed smooth.

We also computed $W^s(p)$ for $a = -0.8, b = 0.1$. Now, $W^s(p)$ intersects J_1 at four points and there are two disjoint segments on $W^s(p)$ that have two pre-images; see Figure 9(b). The primary manifold of $W^s(p)$ is the closed loop that contains p. The two (symmetric) halves of $W^s(p)$ on either side of J_0 account for the two pre-images of the segment in the Z_2 region that contains p in much the same way as described for the case with $a = -0.8$ and $b = 0.2$ in Figure 9(a). The other segment of $W^s(p)$ that lies in the Z_2 region also has two pre-images, but these do not lie on $W^s(p)$. Hence, a disjoint piece containing the pre-images of this segment must exist and it is part of the global stable set $W^s(p)$ of p; see also section 2.1. The set of pre-images forms the disjoint closed loop inside $W^s(p)$ in Figure 9(b), which was also computed with the SC algorithm. To this end, we pick an arbitrary point on the segment of $W^s(p)$ in the Z_2 region that does not contain p. By definition, the Jacobian is nonsingular at this point so that we can use Newton's method to find a pre-image of this point; both are indicated by a red cross in Figure 9(b). We can now apply the SC algorithm to find a new point that maps to $W^s(p)$ and continue in the same way to calculate the entire closed loop inside $W^s(p)$. Again, the images traverse the segment of $W^s(p)$ in the Z_2 region in between the two intersections with J_1 twice; the direction switches each time we cross J_0.

This example shows that it is quite important to know where the J_1 curve lies with respect to the global stable set. Namely, the intersections with J_1 determine whether disjoint pieces exist. In fact, one not only needs to know the curve J_1, but also how many pre-images there are of points in specific regions. Our algorithm is not designed to detect the intersections with J_1, though some information can be obtained by monitoring the direction of the images of successive points p_{k+1} during the computation. In order to find the entire global stable set, one should find the curves J_0 and J_1 and also investigate the number of pre-images in regions separated by J_1. For example, the package Pisces [34] is designed for this purpose, and combining it with the SC algorithm would result in a powerful tool for the study of the global dynamics of noninvertible maps.

4.4. Nien–Wicklin map.

A more complex noninvertible map was investigated in [24] to test the performance of the program Pisces [34] to determine the regions of the phase space with different numbers of pre-images. The map is given as

$$\Lambda \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^4 + \nu x^2 + xy + \mu x \\ (1+a)\, y + b + \epsilon x^2 \end{pmatrix}. \tag{4.3}$$

As in [24], we set $\mu = a = b = \epsilon = 0.1$ and $\nu = -1.9$. The map Λ has a saddle fixed point at $p = (0, -b/a) = (0, -1)$ with a negative stable eigenvalue. The phase space is divided into regions that have different numbers of pre-images by the curve J_1, which is the image of the curve J_0. This map has a more complex folding of type $(Z_0 - Z_2 \ll Z_4)$. The J_1 curve contains two cusps forming a region, called a *dovetail* in [21]; see also Figure 10(c). (This is due to the fact that for these parameter values the map is close to a swallowtail bifurcation in which two cusps are created [2].) Points inside the dovetail have four pre-images, whereas points immediately outside the dovetail have two pre-images. Points above the J_1 curve in the upper left-hand corner have zero pre-images. The authors of [24] compute successive pre-images of p to demonstrate the complexity of this map. The stable set $W^s(p)$ contains all these pre-images, but the stable set itself is not computed in [24].

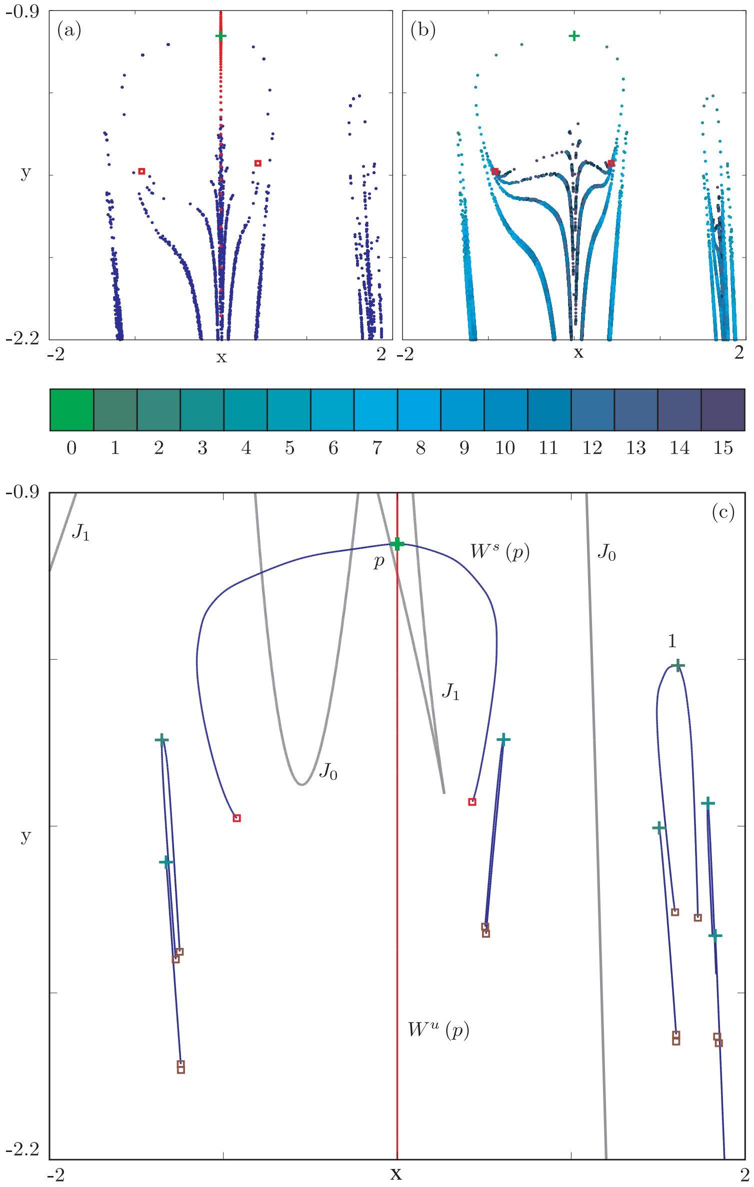


Figure 10. *The stable set (blue) and unstable manifold (red) of the map defined in* (4.3) *with* $\mu = a = b = \epsilon = 0.1$ *and* $\nu = -1.9$*. The red squares in panels* (a), (b) *and the corresponding squares in panel* (c) *indicate a period-two repeller. The other squares in panel* (c) *are pre-images of this period-two repeller. Panel* (a) *shows the stable set computed using the built-in method of DsTool and panel* (b) *shows all points that map onto the saddle after at most* 15 *iterations. The color of the points indicates the number of iterations it takes to map to the saddle point; see the color bar. Panel* (c) *shows the primary manifold computed using the SC method. (The unstable manifold in red is computed up to arclength* 10.*) Note how the primary manifold is extended to the stable set from selected pre-images of the saddle point. All points on the disjoint pieces of the stable set map to the primary manifold within* 12 *iterations.*

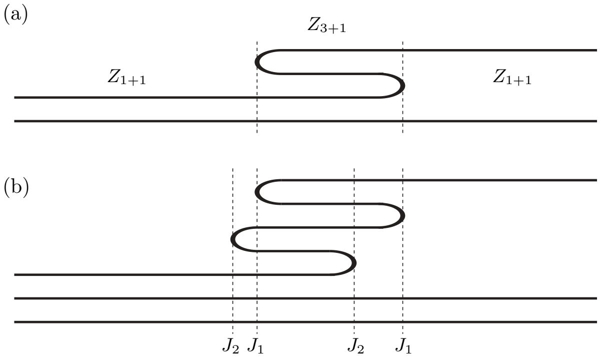


Figure 11. *Illustration of how the plane folds under* Λ (a) *and* Λ^2 (b) *in a neighborhood of the primary manifold. Marked are the intersections with the curves* J_1 *and* J_2 *along which the plane folds.*

We computed $W^s(p)$ with the SC algorithm and the built-in algorithm of DsTool (of fundamental domain iteration) for comparison. Figure 10(a) shows the result of a fundamental domain iteration with DsTool, where 10 uniformly distributed points on a fundamental domain with $\delta = 10^{-6}$ were iterated backward 200 times with Newton's method. Since DsTool uses random seeds, the successive pre-images of p already give an idea of what the stable set looks like. Also shown is the period-two repeller (red squares) that forms the endpoints of the primary manifold of $W^s(p)$. Figure 10(b), by comparison, shows all points that map to the saddle after 15 iterations, as in [24]. These pre-images were obtained with Maple [20] by using the special structure of the map. The number of iterations needed to reach the saddle is indicated by color, ranging from green (1 iterate) to blue (15 iterates). The color coding is indicated in the color bar and is maintained throughout the other figures in this example.

Figure 10(c) shows, in blue, parts of the stable set computed using the SC algorithm. Note that the computations are done using Λ^2 to preserve the orientation. The unstable manifold is shown in red and was computed using the algorithm in [17]; it is the line $\{x = 0\}$. The primary manifold of $W^s(p)$ ends at a period-two repeller and contains a number of pre-images that map directly onto the saddle point. The disjoint parts of the stable set are also computed using the SC algorithm. Indeed, they are grown by starting at pre-images of p indicated by the green crosses in Figure 10(c). The curves J_0 and J_1 are shown in gray. The primary manifold intersects the dovetail formed by the curve J_1; hence, $W^s(p)$ consists of both points with four pre-images and points with two pre-images.

In Figure 11 we illustrate schematically how the plane near $W^s(p)$ folds under Λ. When considering a local neighborhood of $W^s(p)$, the folding of the phase space is what we call type (Z_{1+1}–Z_{3+1}–Z_{1+1}), which is a double fold with another piece of the plane underneath. Since we work with Λ^2 to compute $W^s(p)$, we also illustrate how $W^s(p)$ folds under Λ^2. Now we also have folding along the curve $J_2 = \Lambda(J_1)$ because the critical set of curves along which the plane folds is $J_1(\Lambda^2) = J_1 \cup J_2$, which is the image of $J_0(\Lambda^2) = J_{-1} \cup J_0$, where $\Lambda(J_{-1}) = J_0$; see section 2.1.

Figure 12(a) shows $W^s(p)$ and the curves J_{-1}, J_0, J_1, and J_2. Figure 12(b) gives a schematic representation of $W^s(p)$ and its image under Λ^2. The primary manifold with the

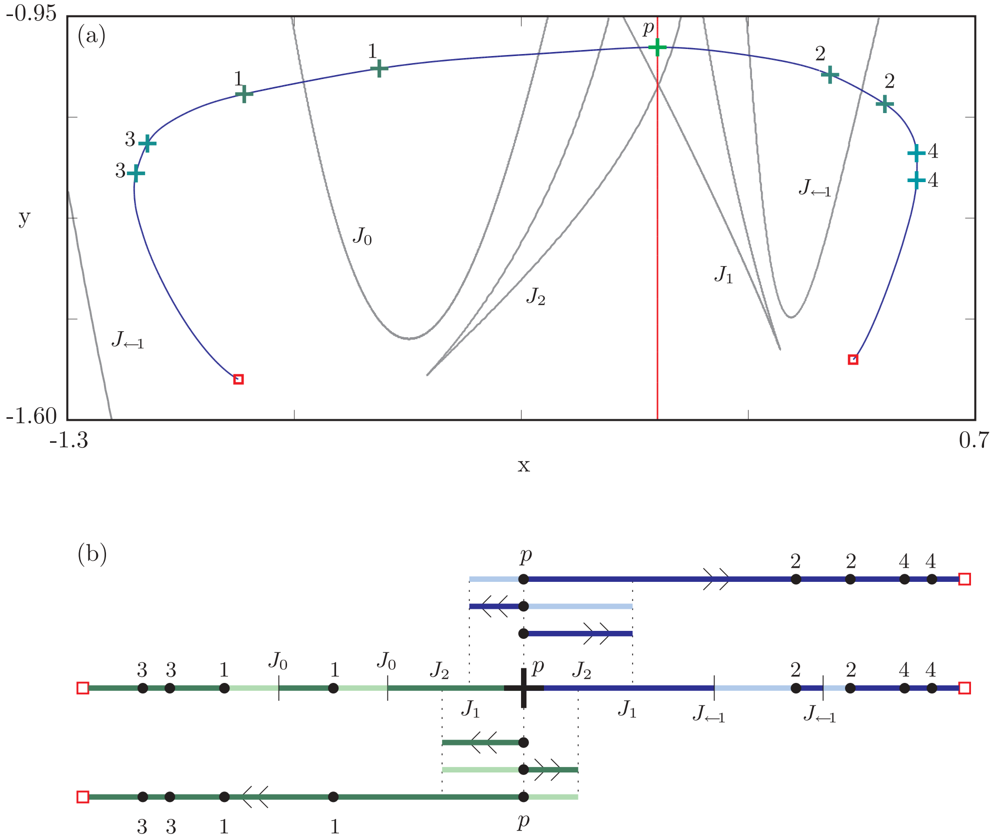


Figure 12. *Panel* (a) *shows the primary manifold along with the the curves* J_{-1}*,* J_0*,* J_1*, and* J_2*. The pre-images of* p *are also indicated by crosses. The color of the crosses is, as in Figure* 10*, indicated by the number of iterations needed for the point to map to the saddle point. Panel* (b) *is a schematic illustration of how the image of* $W^s(p)$ *under* Λ^2 *covers itself. The red squares indicate a period-two repeller. Segments of a particular color map to similarly colored segments, whereby points marked* 3 *and* 4 *are mapped to* 1 *and* 2*, respectively, and* 1 *and* 2 *map to the saddle* p*.*

period-two repellers as its endpoints is shown as a straight line. The saddle point p is the cross in the center. The pre-images of p are marked in the figure by the number of iterations of Λ that it takes to map to p, up to Λ^4. Also marked are the intersections with the curves J_{-1}, J_0, J_1, and J_2. Consecutive sections on the primary manifold are indicated by alternating colors, green for the left branch and blue for the right branch. Correspondingly colored pieces of the Λ^2-image of the primary manifold are shown above and below the primary manifold.

To help interpret this figure, let us start by examining the right-hand side of $W^s(p)$, as was done in section 4.3. The SC algorithm again grows the primary manifold starting from an initial segment with $\delta = 10^{-3}$. The part of the primary manifold from p up to the first intersection with J_{-1} is mapped one-to-one to the segment up to the intersection with J_1, which is indicated by the dark blue piece shown immediately above the branch of the primary manifold. The next segment from the intersection with J_{-1} to the point marked 2, indicated by the light blue, is then mapped back to the piece of branch that traverses back from J_1 to the saddle point p.

The segment between the point marked 2 up until the next point marked 2 maps to the other side of the saddle. Our algorithm stops because we have not yet found this piece of the primary manifold, and so we cannot check if points map onto it. We would have to calculate the branch of the primary manifold on the other side of the fixed point and swap back between the two sides to grow the primary manifold further. For the map Λ, we get around this problem by using the following trick: we try one further iteration of the points to see if they map back to the branch of the primary manifold that was already computed. We can see from Figure 12 that the segment between the two points marked 2 on the branch maps to segments of the branch on the other side of the saddle under Λ^2. The section would map back to the right side of the primary manifold under Λ^3, which is a segment we have already computed.

The segment containing the two points marked 4 maps to the entire right branch under Λ^2. The image of the left branch folds similarly, covering part of the right branch up to $J_2 = \Lambda(J_1)$. Notice how the images of newly found points trace the folded plane under Λ^2, as shown in Figure 11.

In [24] it was observed that the set of pre-images appears to contain teardrop-shaped clusters that are mapped onto each other. We compute other disjoint pieces of $W^s(p)$ by selecting pre-images of the saddle point that lie on these teardrop-shaped pieces. These points are shown in Figure 10(c) as green crosses (color coded relative to the number of iterations it takes to map to p). As in the previous example, we apply the SC algorithm to find a new point that maps to the primary manifold. It is sometimes necessary to change the number of iterates of Λ that are used to map back onto $W^s(p)$.

In Figure 13 we zoom into the primary manifold and a pre-image of it. Figures 13(a) and 13(c) show $W^s(p)$ and a pre-image of $W^s(p)$, computed using Newton's method with the built-in algorithm of DsTool (again, only 10 points on a fundamental domain were iterated backward 200 times). Figures 13(b) and 13(d) show the same pieces of the stable set computed using the SC method. Overlaid on the pieces are pre-images of the fixed point indicated by a colored cross and labeled according to the number of iterations it takes to map back to p. The endpoints of these pieces are pre-images of the period-two repeller, but in themselves they are not period-two points. We can see how pre-images of p on the secondary piece map to pre-images on $W^s(p)$. The disjoint pieces of the stable set are due to pre-images associated with the extra sheet of the plane that lies underneath the folded sheet, as shown in Figure 11.

In Figure 13(d) we labeled the pre-images of the J_0 and J_1 curves as follows. We define $J_0^{-1} \supseteq J_{-1}$ such that $\Lambda(J_0^{-1}) = J_0$, and $J_1^{-1} \supseteq J_0$ such that $\Lambda(J_1^{-1}) = J_1$. This clearly reveals the repetitive structure of the stable set's interaction with the critical curves. Note that there certainly exists a pre-image of the unstable manifold in Figures 13(c) and 13(d), but this pre-image is not part of $W^u(p)$ because the unstable manifold is only forward invariant and is connected to the saddle point.

4.5. The highly interrupted cutting map. We consider a piecewise-smooth map that models a cutting tool. This map was considered in [4, 5, 33] to study the properties of high-speed low immersion milling. The tool oscillates with a fixed frequency of period τ. This

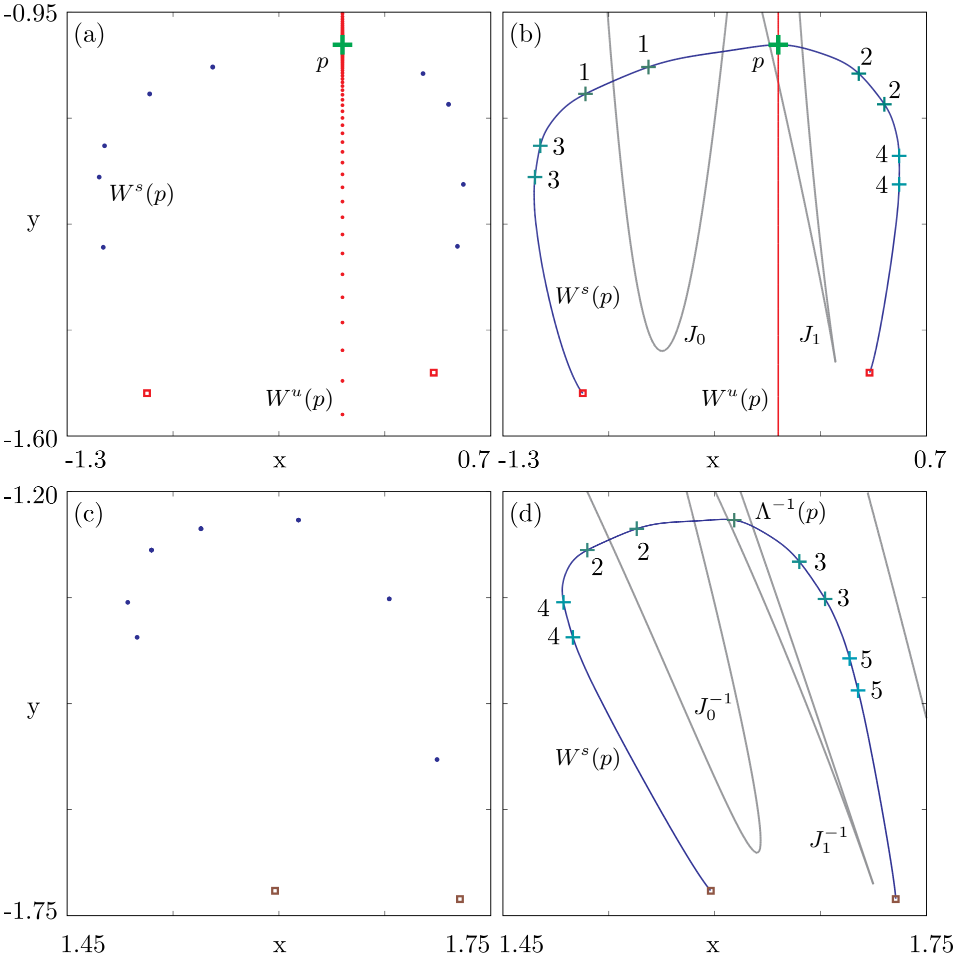


Figure 13. *Two pieces of the stable set of the map* Λ *defined in* (4.3) *with* $\mu = a = b = \epsilon = 0.1$ *and* $\nu = -1.9$. *Panels* (a) *and* (b) *show* $W^s(p)$ *computed using the built-in method of DsTool and the SC algorithm, respectively. Panels* (c) *and* (d) *show a pre-image of* $W^s(p)$ *computed in the same way. The color of the crosses is, as in Figure* 10, *indicated by the number of iterations needed for the point to map to* p.

period is made up of two parts, $\tau_1 + \tau_2 = \tau$. The tool oscillates freely for time τ_1 when the tool is not in contact with the workpiece and the machine re-adjusts its position depending on the thickness of the previous cut. The tool then cuts for time τ_2, which is small compared to τ. The map describes the changes in the height x_j of the cutting tool and its velocity v_j at the moment just after the cutting time τ_2. The map is defined, as given in [33], by

$$(4.4) \qquad \begin{cases} x\left(t_{j+1}\right) = A_{11}x\left(t_j\right) + A_{12}v\left(t_j\right), \\ v\left(t_{j+1}\right) = A_{21}x\left(t_j\right) + A_{22}v\left(t_j\right) + F(x(t_j), (v(t_j)), \end{cases}$$

where A_{11}, A_{12}, A_{21} and A_{22} are functions of the period τ and the relative damping factor ζ, namely,

$$
\begin{aligned}
A_{11} &= \tfrac{e^{-\zeta\tau}}{(1-\zeta^2)}\left(\cos\left(\sqrt{(1-\zeta^2)}\tau\right)+\zeta\sin\left(\sqrt{(1-\zeta^2)}\tau\right)\right),\\
A_{12} &= \tfrac{e^{-\zeta\tau}}{(1-\zeta^2)}\sin\left(\sqrt{(1-\zeta^2)}\tau\right),\\
A_{21} &= -\tfrac{e^{-\zeta\tau}}{(1-\zeta^2)}\sin\left(\sqrt{(1-\zeta^2)}\tau\right),\\
A_{22} &= \tfrac{e^{-\zeta\tau}}{(1-\zeta^2)}\left(\cos\left(\sqrt{(1-\zeta^2)}\tau\right)-\zeta\sin\left(\sqrt{(1-\zeta^2)}\tau\right)\right).
\end{aligned}
$$

The function F in the equation for $v(t_{j+1})$ is defined as

$$
F(x(t_j), v(t_j)) = \begin{cases} \tfrac{\tau_2}{m\omega_n^s} F_c\left(h\right) & \text{if } h = h_0 + x\left(t_j\right) - A_{11}x\left(t_j\right) - A_{12}v\left(t_j\right) > 0\,,\\ 0 & \text{otherwise.} \end{cases}
$$

Here, m is the mass of the tool, h_0 is the feed during one cutting period, and ω_n is the natural damping frequency. The cutting force $F_c(h)$ is estimated by the *three-quarters rule*, an empirical formula, giving $F_c(h) = Kwh^{3/4}$ for some parameter K, where w is the chip width and h defines the cutting depth. This term is only included for $h = h_0 - A_{11}x\left(t_j\right) - A_{12}v(t_j) + x\left(t_j\right) > 0$, that is, when the tool cuts the material, and is zero otherwise; see [4, 5, 33] for more details on the derivation of this model. We combine the constants τ_2, m, ω_n, w and the parameter K, giving a new parameter K_2. The equation $h = 0$ defines a straight line in the (x, v)-space along which the system is not differentiable. The map (4.4) is complicated and finding an explicit inverse is impractical.

The fixed point is not on the smoothness boundary $\{h = 0\}$, so in a neighborhood of the saddle the map is smooth and local manifolds exist. Globally, the system is piecewise smooth and we refer to the global manifolds as *piecewise-smooth manifolds*.

Figure 14 shows the piecewise-smooth manifolds of the map defined in (4.4). For our computation we chose $\zeta = 0.01$, $\tau = 2$, and $h_0 = 1$. For $K_2 = 0.87$ there is a saddle point p at $x_e = 0.27932$, $v_e = 0.43835$, with two negative eigenvalues. This is indicated by a green cross in the figure. Figure 14(a) shows the two branches of the stable manifold in blue computed up to arclength 400 using the SC method for the second iterate of the map. The unstable manifold, shown in red, is computed up to arclength 10 using the algorithm from [17].

The branches of the stable manifold spiral outward with obvious nonsmooth corners. The branches of the unstable manifold also have nonsmooth corners as they approach the period-two repeller; see the enlargement in Figure 14(b). The gray line is the smoothness boundary (switching manifold) along which $h = 0$.

The global stable manifold is nonsmooth at the point where it intersects the smoothness boundary, but there are also nonsmooth points that are not on this boundary. These nonsmooth points are pre-images of intersections with the smoothness boundary, so the nonsmoothness propagates backward along the manifold. Hence, the stable manifold is smooth until it arrives at the smoothness boundary or a pre-image of the smoothness boundary. To illustrate this, some of the intersections with the boundary are marked by a colored cross and a zero in Figure 14. Points that map to these intersection points are indicated by similarly colored crosses labeled with a number indicating the number of iterations it takes to map to the first intersection point.

Figure 14(b) shows a close-up of the piecewise-smooth unstable manifold. The unstable manifold was computed with the algorithm in [17] and accumulates on a period-two attractor.

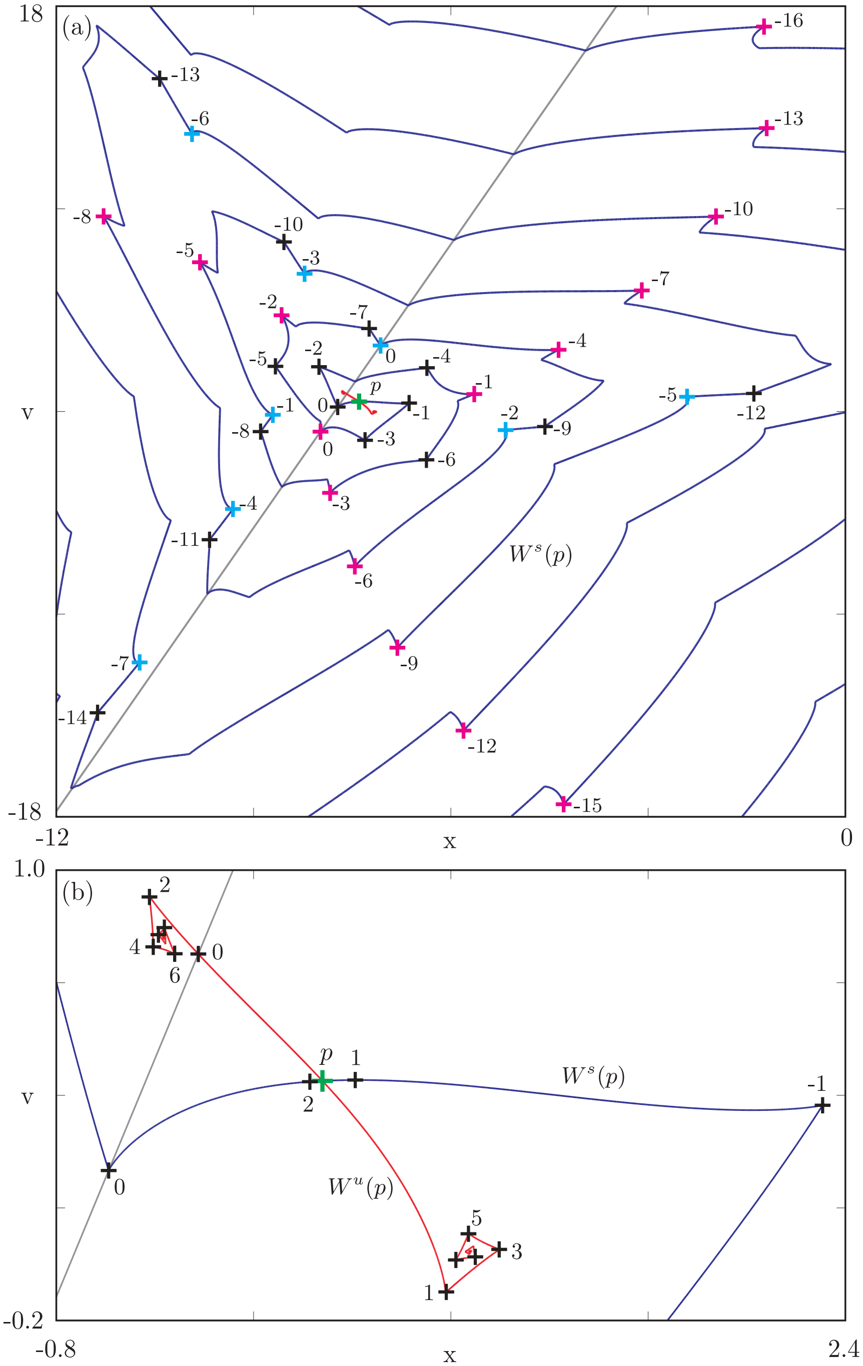


Figure 14. *Panel* (a) *shows the stable (blue) and unstable (red) piecewise-smooth manifolds of the saddle p of the map defined in* (4.4) *for $\zeta = 0.01$, $\tau = 2$, $h_0 = 1$, $K_2 = 0.87$. We computed the stable manifold up to arclength* 400 *and the unstable manifold until it converged to the period-two points. The gray line indicates the smoothness boundary. Crosses with the same color are part of the same orbit on the stable manifold, which indicates how sharp corners on the manifold map to intersections with the smoothness boundary. An enlargement of $W^u(p)$ is shown in panel* (b)*; the black crosses on the unstable manifold are images of the intersection point with the smoothness boundary. This clearly shows how sharp corners emanate from the smoothness boundary.*

We see that the intersection of the unstable manifold with the smoothness boundary is smooth, but sharp corners propagate forward along the manifold from this intersection point.

In [33] the authors perform a bifurcation analysis on the parameter K. We can similarly vary the parameter K_2. For $K_2^* = 0.856468$ there is a subcritical period-doubling bifurcation. Here, the fixed point becomes stable and a period-two saddle emerges between the fixed point and the period-two attractor. In contrast to the period-two attractor, both points on the period-two saddle lie on the same side of the smoothness boundary. In Figure 15 we computed the stable manifolds (blue) and the unstable manifolds (red) of the two period-two saddle points for $K_2 = 0.85$. The period-two saddles are indicated by the green crosses and the attractors by blue triangles. The piecewise-smooth stable manifolds produced using the SC algorithm in this case define the boundaries of the basins of attraction between the period-two attractor and the attracting fixed point. Points in the white region tend to the attracting fixed point, whereas points in the shaded regions oscillate between the blue and gray regions while tending to the period-two attractor.

5. Conclusions and discussion.

Many real-life systems are very hard to invert, and thus the inverse is not known explicitly, or is truly noninvertible with several branches of inverses. The SC algorithm described in this paper computes the one-dimensional global stable manifold, or stable set, for a planar map without requiring knowledge of the inverse map or the Jacobian matrix. Even if the inverse map is known explicitly, the SC algorithm can still be used to compute the stable manifold.

Until now, stable sets of systems with multiple pre-images could not be easily computed past points where the Jacobian is singular. We have shown that the SC algorithm is not affected by this problem. Furthermore, the SC algorithm can also be used to compute disjoint pieces of the stable set, as soon as one point on it is found. Such a point is a pre-image of an arbitrary point on the already computed manifold, for example, the fixed point. Finally, we have shown that our algorithm also works for maps in which the derivative of the system is not continuous. To our knowledge, manifolds of such piecewise-smooth systems have not previously been computed. We believe that these properties of the algorithm are very useful when one is exploring real-life systems.

The SC algorithm has been implemented to compute one-dimensional stable manifolds and stable sets of planar maps. Although planar maps (already) display many types of dynamical behavior, including chaos, interesting dynamics can also be investigated in higher dimensions. Let us discuss the implications of generalizing the SC algorithm to higher dimensions. To compute a two-dimensional stable manifold in a three-dimensional phase space, we could use a direct combination of [18] and the SC algorithm. In fact, any system with codimension-one stable manifolds can be computed by generalizing the method in [18] and using it in combination with the SC algorithm. The complexity of the algorithm is of the same order as the method in [18] for dimensions greater than two.

As mentioned in the introduction, the PIM-triple procedure of [25], designed for finding chaotic saddles, can be used to calculate stable sets of saddle points. The PIM-simplex method in [23] is an extension of the PIM-triple method in [25] for computing chaotic saddles having an unstable dimension of two or higher. This method uses a nondegenerate simplex to probe the space replacing the line segment; see also [32]. Our algorithm could be generalized in a

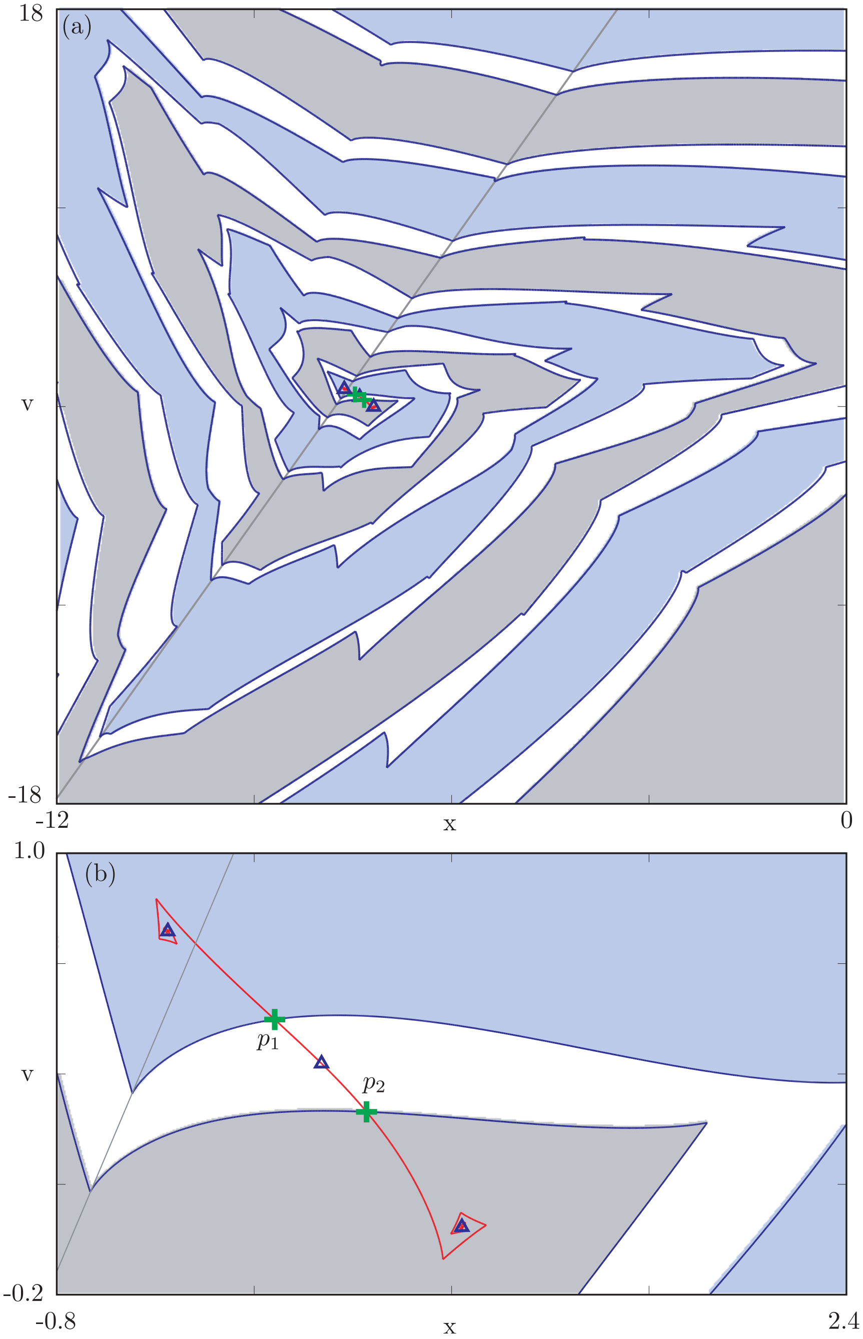


Figure 15. *The stable (blue) and unstable (red) manifolds* (a) *of a period-two saddle of the map defined in* (4.4) *for* $\zeta = 0.01$, $\tau = 2$, $h_0 = 1$, $K_2 = 0.85$. *Again, the stable manifolds are computed up to arclength* 400 *and the unstable manifolds up to arclength* 10. *The gray line indicates the smoothness boundary. An enlargement of* $W^u(p)$ *is shown in panel* (b).

similar way. Instead of searching over a segment of a circle in two dimensions for the next point, we could search over the surface of an n-dimensional hypersphere in n dimensions. However, computing only one-dimensional manifolds of an n-dimensional system would lead to an increased complexity that is not reflected in the dimension of the manifold. For these cases, knowledge of the inverse would dramatically speed up the computation time.

We expect the SC algorithm to be useful in helping to understand the dynamics of many real-life noninvertible and/or nonsmooth systems. The great advantage of this method is that only a minimal amount of knowledge of the system is required to compute the stable manifold or the stable set.

Acknowledgments. We thank Tim Burns, Gábor Stépán, and Róbert Szalai for helpful discussions and assistance with the highly interrupted cutting map. We are grateful to Bruce Peckham for pointing out the example of a global stable set that forms a closed loop.

REFERENCES

[1] R. H. Abraham, L. Gardini, and C. Mira, *Chaos in Discrete Dynamical Systems, A visual introduction in* 2 *Dimensions*, Springer-Verlag, New York, 1997.

[2] V. I. Arnol′d, *Catastrophe Theory*, 3rd ed., Springer-Verlag, Berlin, 1992.

[3] A. Back, J. Guckenheimer, M. R. Myers, F. J. Wicklin, and P. A. Worfolk, *DsTool: Computer assisted exploration of dynamical systems*, Notices Amer. Math. Soc., 39 (1992), pp. 303–309.

[4] M. A. Davis, J. R. Pratt, B. Dutterer, and T. J. Burns, *The stability of low radial immersion*, Ann. CIRP, 49 (2000), pp. 37–40.

[5] M. A. Davis, J. R. Pratt, B. Dutterer, and T. J. Burns, *Stability prediction for low radial immersion milling*, J. Manufacturing Sci. Engrg., 124 (2002), pp. 217–225.

[6] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.

[7] C. E. Frouzakis, L. Gardini, I. G. Kevrekidis, G. Millerioux, and C. Mira, *On some properties of invariant sets of two-dimensional noninvertible maps*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 7 (1997), pp. 1167–1194.

[8] C. E. Frouzakis, I. G. Kevrekidis, and B. Peckham, *A route to computational chaos revisited: Noninvertibility and the breakup of an invariant circle*, Phys. D 177 (2003), pp. 101–121.

[9] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, Berlin, 1983.

[10] I. Gumowski and C. Mira, *Dynamique Chaotique*, Cepadues Éditions, Toulouse, 1980.

[11] I. Gumowski and C. Mira, *Recurrences and Discrete Dynamic Systems*, Springer-Verlag, New York, 1980.

[12] S. M. Hammel, C. K. R. T. Jones, and J. V. Maloney, *Global dynamical behaviour of the optical field in a ring cavity*, J. Opt. Soc. Amer. Ser. B, 2 (1985), pp. 552–564.

[13] D. Hobson, *An efficient method for computing invariant manifolds*, J. Comput. Phys., 104 (1991), pp. 14–22.

[14] A. J. Homburg, H. M. Osinga, and G. Vegter, *On the computation of invariant manifolds of fixed points*, Z. Angew. Math. Phys., 46 (1995), pp. 171–187.

[15] K. Ikeda, *Multiple-valued stationary state and its instability of the transmitted light from a ring cavity system*, Optics Communications, 30 (1979), pp. 257–261.

[16] E. J. Kostelich, J. A. Yorke, and Z. You, *Plotting stable manifolds: Error estimates and noninvertible maps*, Phys. D, 93 (1996), pp. 210–222.

[17] B. Krauskopf and H. M. Osinga, *Growing* 1*D and quasi-*2*D unstable manifolds of maps*, J. Comput. Phys., 146 (1998), pp. 406–419.

[18] 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.

[19] B. Krauskopf and H. M. Osinga, *Investigating torus bifurcations in the forced Van der Pol oscillator*, in Numerical Methods for Bifurcation Problems and Large-Scale Dynamical Systems, E. J. Doedel and L. S. Tuckerman, eds., IMA Vol. Math. Appl. 119, Springer-Verlag, New York, 2000, pp. 199–208.

[20] Maplesoft Product Suite, http://www.maplesoft.com/ (2004).

[21] C. Mira, C. Jean-Pierre, G. Millérioux, and L. Gardini, *Plane foliation of two-dimensional noninvertible maps*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 6 (1996), pp. 1439–1462.

[22] C. Mira, L. Gardini, A. Barugola, and J. C. Cathala, *Chaotic Dynamics in Two-Dimensional Noninvertible Maps*, World Scientific Ser. Nonlinear Sci. Ser A: Monographs and Treaties 20 A, World Scientific, River Edge, NJ, 1996.

[23] P. Moresco and S. P. Dawson, *The PIM-simplex method: An extension of the PIM-triple method to saddles with an arbitrary number of expanding directions*, Phys. D, 126 (1999), pp. 38–48.

[24] C.-H. Nien and F. J. Wicklin, *An algorithm for the computation of preimages in noninvertible mappings*, 8 (1998), pp. 415–422.

[25] H. E. Nusse and J. A. Yorke, *A procedure for finding numerical trajectories in chaotic saddles*, Phys. D, 36 (1989), pp. 137–156.

[26] H. M. Osinga and J. P. England, *Global Manifold* 1*D Code, Version* 2, Software for Use with DsTool, http://www.dynamicalsystems.org/sw/sw/detail?item=27 (2003).

[27] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems*, Springer-Verlag, New York, Berlin, 1982.

[28] J. Palis and F. Takens, *Hyperbolicity and Sensitive Chaotic Dynamics at Homoclinic Bifurcations*, Cambridge University Press, Cambridge, UK, 1993.

[29] T. S. Parker and L. O. Chua, *Practical Numerical Algorithms for Chaotic Systems*, Springer-Verlag, Berlin, 1989.

[30] C. Simó, *On the analytical and numerical approximation of invariant manifolds*, in Les Méthodes Modernes de la Mécanique Céleste, D. Benest and C. Froeschlé, eds., Goutelas, 1989, pp. 285–330.

[31] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*, Perseus Books, Reading, MA, Cambridge, MA, 1994.

[32] D. Sweet, H. E. Nusse, and J. A. Yorke, *Stagger-and-step method: Detecting and computing chaotic saddles in higher dimensions*, Phys. Rev. Lett., 86 (2001), pp. 2261–2264.

[33] R. Szalai, *Nonlinear Vibrations of Interrupted Cutting Processes*, M.Sc. thesis, Budapest University of Technology and Economics, Budapest, Hungary, 2002.

[34] F. J. Wicklin, *Pisces: A Platform for Implicit Surfaces and Curves and the Exploration of Singularities*, Tech. report GCG 89, The Geometry Center, University of Minnesota, Minneapolis, MN, 1995. Available online at http://www.geom.uiuc.edu/~fjw/pisces/.

[35] Z. You, E. J. Kostelich, and J. A. Yorke, *Calculating stable and unstable manifolds*, 1 (1991), pp. 605–623.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 3, pp. 191–231

Pattern Formation in a Network of Excitatory and Inhibitory Cells with Adaptation[*]

Rodica Curtu[†] and Bard Ermentrout[†]

Abstract. A bifurcation analysis of a simplified model for excitatory and inhibitory dynamics is presented. Excitatory cells are endowed with a slow negative feedback and inhibitory cells are assumed to act instantly. This results in a generalization of the Hansel–Sompolinsky model for orientation selectivity. Normal forms are computed for the Turing–Hopf instability, where a new class of solutions is found. The transition from stationary patterns to traveling waves is analyzed by deriving the normal form for a Takens–Bogdanov bifurcation. Comparisons between the normal forms and numerical solutions of the full model are presented.

Key words. pattern formation, neural networks, normal forms, Takens–Bogdanov

AMS subject classifications. 34C15, 34C23, 34C25, 37G15, 37N25, 92C20

DOI. 10.1137/030600503

1. Introduction.

Many areas of the brain have a sheet-like architecture in which neurons interact horizontally through some spatial distribution. This is useful since it provides a substrate for topographic connectivity in sensory areas such as vision and touch. For example, in vision, nearby areas in the retina project to nearby areas in the visual cortex [16]. Spatial coding is not the only organizing principle in cortical networks. Neurons that respond preferentially to similar stimuli are more strongly linked than those responding to different stimuli. A classic example of this is orientation preference in the visual cortex [15]. Neurons in parts of the visual cortex respond to oriented bars of light and many of these cells have a preferred angle. Thus, topologically, we can regard the network as lying on a ring of length π corresponding to all possible orientations. Similar organizing principles based on angular preference are also seen in the head-direction system of rodents [25]. Experiments in which slices are removed from the cortex and pharmacologically manipulated (to make them more excited) support propagating waves [2]. This shows that the "wiring" between neurons is spatially organized.

In the simplest sense, cortical neurons can be divided into excitatory and inhibitory cells. Most excitatory, or regular-spiking, neurons have some form of spike-frequency adaptation (SFA). That is, there is a slow intrinsic negative feedback term which lowers the firing rate of the neuron when a constant current is applied. In contrast, inhibitory, or fast-spiking, neurons generally do not have a similar rate adaptation. In the classic models of Wilson and Cowan [24], excitatory and inhibitory cells are treated the same with a single scalar number (the firing rate) associated with each neuron. In a recent model for feature selectivity, Hansel

[*] Received by the editors July 18, 2003; accepted for publication (in revised form) by M. Golubitsky February 4, 2004; published electronically July 6, 2004. This work was supported in part by the National Science Foundation.
http://www.siam.org/journals/siads/3-3/60050.html

[†] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (rodicacurtu@unitbv.ro, bard@pitt.edu).

and Sompolinsky [15] introduce a rate model in which the excitatory cells are endowed with SFA. The domain of their model is a circle corresponding to orientation preference. Because the model is piecewise-linear they are able to analyze a number of interesting patterns. The motivation for the model is to shed light on the importance of recurrent (nonlinear feedback) connections in setting the orientation selectivity to broadly tuned inputs. (That is, they present an input with a small local maximum in orientation and ask if the network can amplify the difference between the peak and the background.) They find that, even in the absence of input, their network is able to support local peaks of activity; the inputs serve to position the peak by breaking the rotational symmetry of the model. Interestingly, they show that the presence of SFA causes the local peak to rotate around the circle in the form of a traveling wave. They exploit the piecewise linear nature of their model to determine parameter regimes where traveling waves exist as a function of the adaptation.

Our goal in this paper is to generalize the nonlinearity used in their model and then use methods from dynamical systems to study and characterize the nonlinear behavior. We first state the model equations and simplify it so that it becomes a two-variable model (neural firing rate and degree of adaptation). The spatial interactions are of the lateral-inhibition type; there is local excitation of nearby spatial points (or features) and inhibition at distant points. With weak or rapidly decaying adaptation, we show that there is a bifurcation to stationary spatial patterns. These correspond to the peaked patterns computed in the Hansel–Sompolinsky (HS) model. Moreover, with strong or slowly decaying adaptation, there is a Hopf bifurcation to spatially varying patterns. Under some circumstances, these correspond to the traveling waves in the HS model; however, depending on the nonlinearity, there are also standing oscillations. Finally, we look at the point of transition between stationary and oscillating patterns. This occurs at a Takens–Bogdanov bifurcation and leads to some interesting and complex behavior. This latter part occupies the bulk of the paper.

While normal forms have been computed generally for systems undergoing a Hopf bifurcation or Takens–Bogdanov bifurcation in the presence of $O(2)$ symmetry, there are many qualitatively different normal forms. Thus, in order to apply these methods to a specific problem, it is necessary to actually compute the coefficients. Thus, we describe the highlights of the calculation in the main text and leave the details to Appendix A. One of the main reasons for computing the coefficients is so that we can find regions in parameter space, where new types of behavior occur. In particular, we have found conditions on the nonlinearity such that standing waves occur. These are not found in the original HS model, where the nonlinearity was piecewise-linear.

1.1. HS model. Hansel and Sompolinsky [15] introduced a simple rate-model for the study of feature selectivity in local cortical circuits. In that context the network of neurons was assumed to code for a sensory or movement scalar feature x (for example, the angle a bar is rotated in the subject receptor field so that x can be taken in the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$). The local cortical network consists of ensembles of neurons that respond (are tuned) to a particular feature of an external stimulus, and so are called "feature columns," and that are interconnected by recurrent synaptic connections. In other words, each neuron in the network is selective, firing maximally when a feature ("preferred feature" of the neuron) with a particular value is present. The synaptic interactions between a presynaptic neuron y from the β-population and a postsynaptic neuron x from the α-population are denoted by a function

$J^{\alpha\beta}(x-y) = j_0^{\alpha\beta} + j_2^{\alpha\beta}\cos(2(x-y))$, where α and β indices stand for E (excitatory) and/or I (inhibitory) populations of neurons, depending on the context. We take $j_0^{\alpha E} \geq j_2^{\alpha E} \geq 0$ for input coming from the excitatory population and $j_0^{\alpha I} \leq j_2^{\alpha I} \leq 0$ for input coming from the inhibitory population.

Hansel and Sompolinsky collapsed both excitatory and inhibitory populations into a single equivalent population. In this case the synaptic connectivity function J is defined as $J(x-y) = j_0 + j_2\cos(2(x-y))$ with no restrictions on the sign of coefficients, and the rate-model has a single rate variable $m(x,t)$ that represents the activity of the population of neurons in the column x at time t. Moreover, the population is assumed to display adaptation. The resulting model [15] is

$$\tau_0 \frac{\partial m}{\partial t}(x,t) = -m(x,t) + F\left(\frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} J(x-y)m(y,t)dy + I^0(x-x_0) - I_a(x,t) - T\right),$$

$$\tau_a \frac{\partial I_a}{\partial t}(x,t) = -I_a(x,t) + J_a\, m(x,t)\,. \tag{1.1}$$

I^0 stands for the synaptic currents from the external neurons, T is the neuronal threshold, I_a is the adaptation current, $\tau_a > \tau_0$ is its time constant, and J_a measures the strength of adaptation.

An additional assumption for (1.1) is that the stable state of the network is such that all the neurons are far from their saturation level, allowing the gain function F to be in a semilinear form $F(I) = I$ for $I > 0$, and zero otherwise.

The model we analyze in this paper is based on the above HS model but includes a more general nonlinear gain function F. It also can be used in a more general context of synaptically coupled populations of excitatory and inhibitory neurons with adaptation.

1.1.1. A more general class of models. More generally, the problem we are interested in concerns the possible patterns that can be obtained in a neuronal network consisting of both excitatory and inhibitory cells, and in the presence of adaptation. The network model consists of two homogeneous populations of neurons, one excitatory (E), displaying adaptation, and the other one inhibitory (I), without adaptation. This is a reasonable assumption, for example, for *cortical neurons*, since experimental studies report that in the cortex most of the inhibitory neurons do not display spike adaptation [3], [4], [14].

The spatial connectivity is assumed to be all-to-all from E to E, E to I, and I to E cells, and in all cases the strength of the interactions decreases with the distance between neurons according to a Gaussian distribution with zero mean, say J_{EE}, J_{IE}, and J_{EI}, respectively. For simplicity, no I to I interactions are included. In addition, the network is considered one-dimensional in space.

We assume in the following a *linear adaptation* and describe the neuronal activity by a rate-model. That is, we have

$$\tau_E \frac{du_E}{dt} = -u_E + F_E(J_{EE} * u_E - J_{EI} * u_I - gA)\,,$$

$$\tau_I \frac{du_I}{dt} = -u_I + F_I(J_{IE} * u_E)\,,$$

$$\tau_A \frac{dA}{dt} = -A + u_E\,, \tag{1.2}$$

where τ_E, τ_I, τ_A are the time constants for the excitatory and inhibitory neurons, and for adaptation, respectively; A is the variable that defines the adaptation; g is the strength of adaptation; F_E and F_I are the firing-rate functions; and $J_{ij} * u_j$, with $i, j \in \{E, I\}$, is the convolution $J_{ij} * u_j(x,t) = \int_{-\infty}^{\infty} J_{ij}(x-y) u_j(y,t)\, dy$. Pinto and Ermentrout [20] analyzed a model like this, without the inhibitory interactions, in order to study propagating waves.

One simplification is to assume that the inhibition is much faster than the excitation, and that the firing rate for the inhibitory population is linear. That allows us to replace the equation for I cells with its steady state, i.e., to take $u_I \approx F_I(J_{IE} * u_E) = J_{IE} * u_E$. Then, since a convolution of two Gaussians with zero mean is still a Gaussian with zero mean, we have $(J_{EE} * u_E - J_{EI} * u_I)(x,t) = (\, J_{EE} - J_{EI} * J_{IE}\,) * u_E(x,t) = J * u_E(x,t)$, where $J(x)$ is a difference of two Gaussians. Therefore system (1.2) can be reduced to a rate-model for only one variable u, in which we include the neuronal activity for both excitatory and inhibitory populations, and with the synaptic coupling defined by a function J as in Figure 1(a) (the “Mexican hat”).

In [10] one of us showed the existence of traveling waves in a two-population model without adaptation. In this work, in order to get a Hopf bifurcation at a nonzero wavenumber, it was necessary to assume that the inhibition was slow and that the range of inhibitory-inhibitory interactions exceeded that of the excitatory-excitatory interactions. This is not a realistic assumption for the sensory cortex, where the inhibition is more localized and acts rapidly. In this older work, excitatory cells were treated the same as inhibitory cells. However, it is now known that excitatory cells exhibit pronounced adaptation; thus we have included this in the model. Below, we describe conditions under which both Hopf and Takens–Bogdanov bifurcations occur in the more general problem, where the inhibitory cells are treated as a separate population.

1.1.2. Mathematical model.

Under the assumptions considered in the previous section, the mathematical model equivalent to (1.2) is

$$
\begin{aligned}
\frac{\partial u}{\partial t} &= -u(x,t) + F\left(\alpha\, J * u\,(x,t) - g\, v(x,t)\right), \\
\tau\, \frac{\partial v}{\partial t} &= -v(x,t) + u(x,t)
\end{aligned}
\tag{1.3}
$$

with $x \in \mathbb{R}$ the one-dimensional spatial coordinate, and α, g, and τ positive parameters.

The variables u and v represent the neuronal activity and adaptation, respectively, τ and g correspond to the time constant and the strength of adaptation, and α is a parameter that controls the strength of the synaptic coupling J.

Synaptic coupling. J is a continuous and even function, $J(-x) = J(x)\ \forall x \in \mathbb{R}$, and is absolutely integrable on the interval $[-l, l]$, where $l \in \mathbb{R}_+ \cup \{\infty\}$. If $l = \infty$, we ask that $\lim_{x\to-\infty} J(x) = \lim_{x\to\infty} J(x) = 0$. Otherwise, J is assumed to be periodic of period $2l$. Then the operator $J * u$ is defined as

$$
J * u\,(x,t) = \int_{-l}^{l} J(x-y)\, u(y,t)\, dy\,. \tag{1.4}
$$

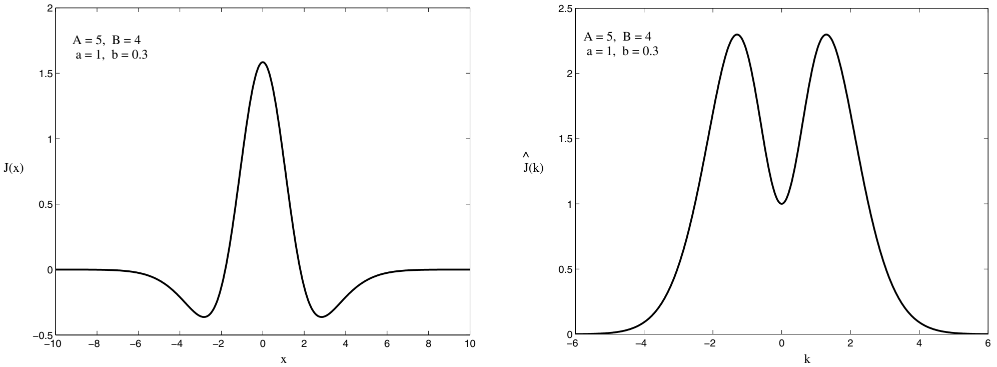


Figure 1. (a) *The coupling* $J(x)=\frac{1}{\sqrt{\pi}}\big[A\sqrt{a}\,e^{-ax^2}-B\sqrt{b}\,e^{-bx^2}\big]$ *for* $A=5$, $B=4$, $a=1$, $b=0.3$. (b) *The function* $\hat{J}$.

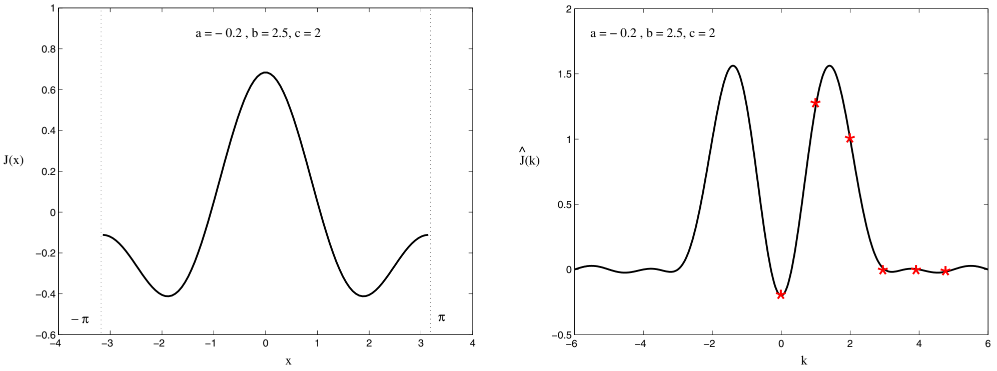


Figure 2. (a) *The periodic coupling* $J(x)=\frac{1}{2l}\big[\,a+b\cos(\frac{\pi x}{l})+c\cos(\frac{2\pi x}{l})\,\big]$ *for* $l=\pi$ *and* $a=-0.2$, $b=2.5$, $c=2$. (b) *The function* $\hat{J}$.

There is an operator associated to J, which is defined on the frequency space, and that is

$$\hat{J}(k)=\int_{-l}^{l} J(x)\,e^{ikx}\,dx\,. \tag{1.5}$$

Remark 1. If we consider an infinite neural network, we take $l=\infty$ and the function J is typically as in Figure 1(a). For example, we can define J as

$$J(x)=\frac{1}{\sqrt{\pi}}\left[A\sqrt{a}\,e^{-ax^2}-B\sqrt{b}\,e^{-bx^2}\right],\ x\in\mathbb{R}\,, \tag{1.6}$$

where $A\geq B>0$, $a>b>0$. Then $\hat{J}(k)=A\,e^{-k^2/4a}-B\,e^{-k^2/4b}$, $k\in\mathbb{R}$, and $\hat{J}$ has the graph as in Figure 1(b). Nevertheless, in numerical simulations we cannot consider an infinite domain. Therefore we have to restrict ourselves to a finite domain $[-l,l]$ with $l\in\mathbb{R}_+$ and work with periodic boundary conditions. In order to maintain the assumptions of local excitation and long range inhibition, J is typically as in Figure 2(a). For example, we can take J as

$$J(x)=\frac{1}{2l}\left[a+b\,\cos\left(\frac{\pi x}{l}\right)+c\,\cos\left(\frac{2\pi x}{l}\right)\right],\ x\in\mathbb{R}\,, \tag{1.7}$$

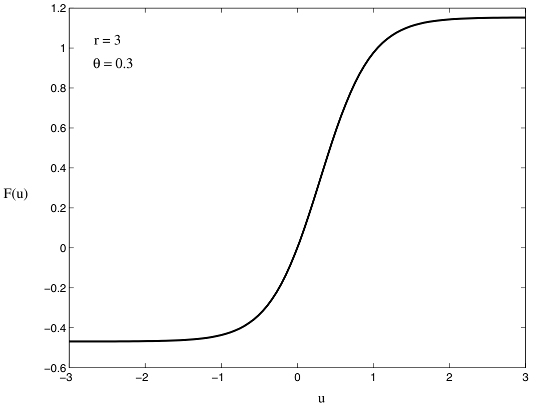


Figure 3. *The firing rate* (1.8) *for* $r = 3$ *and* $\theta = 0.3$.

where a, b, c are real parameters. Therefore $\hat{J}$ (see Figure 2(b)) is

$$\hat{J}(0) = a\,,\;\; \hat{J}(\pm\pi/l) = b/2\,,\;\; \hat{J}(\pm 2\pi/l) = c/2\,,\;\; \hat{J}(\pm j\pi/l) = 0 \;\; (j \in \mathbb{N} \setminus \{0,1,2\})\,,$$

$$\hat{J}(k) = \frac{\sin(lk)\,[(a-b+c)(lk/\pi)^4 + (-5a+4b-c)(lk/\pi)^2 + 4a]}{lk\,[\,(lk/\pi)^2 - 1\,][\,(lk/\pi)^2 - 4\,]}\,,\; k \notin \pm(\pi/l)\mathbb{N}\,.$$

Firing rate. F in (1.3) is a sigmoid function (Figure 3) assumed to satisfy

$$F(0) = 0\,,\;\; F'(0) = 1\,.$$

The first condition translates the steady state to the origin $\bar{u} = 0$, $\bar{v} = 0$. The second condition brings additional simplifications to our calculations. A typical expression for F is then $F(u) = K\big[\frac{1}{1+e^{-r(u-\theta)}} - \frac{1}{1+e^{r\theta}}\big]$, with r and θ positive parameters, and $K = (1+e^{r\theta})^2\, e^{-r\theta}/r$, i.e.,

$$F(u) = \frac{1+e^{r\theta}}{r} \cdot \frac{1-e^{-ru}}{1+e^{-r(u-\theta)}}\,. \tag{1.8}$$

Remark 2. The condition $F'(0) = 1$ is not essential. As long as $F'(0)$ is nonzero and positive, the results proved in the following sections remain valid. To see this, let us assume that $F'(0) \neq 1$. Then, by the change of variables $u_{\text{new}} = u/F'(0)$, $v_{\text{new}} = v/F'(0)$, the change of parameters $\alpha_{\text{new}} = F'(0)\,\alpha$, $g_{\text{new}} = F'(0)\,g$, and the change of function $F_{\text{new}} = F/F'(0)$, we obtain a system topologically equivalent to (1.3), where F_{new} satisfies the constraints $F_{\text{new}}(0) = 0$ and $F^{'}_{\text{new}}(0) = 1$.

Remark 3. The shape of F is also not crucial since we need only local properties of F, such as its first few derivatives. Thus, all we need for the analysis is that F have continuous third derivatives at the origin. However, most neural models use some sort of sigmoidal nonlinearity. The piecewise-linear model is dangerous since solutions can grow without bound under some circumstances.

2. Linear stability analysis and pattern initiation mechanism. Previous studies on reaction diffusion pattern generation mechanisms (see [19] for a review) and neural models of pattern generation (such as a mechanism for stripe formation in the visual cortex [22], a model for the brain mechanism underlying visual hallucination patterns [10], [12], or a neural activity model for shell patterns [11]) indicate that in one-dimensional structures the linear theory turns out to be a good predictor of the ultimate steady state of the full nonlinear system. There is very good agreement between the theoretical solutions obtained from the linearized problem, and the numerical simulations of the original nonlinear system with initial conditions taken to be small random perturbations about the steady state.

Nevertheless, in order to find the solution of the linearized problem that corresponds to the stable spatial or spatio-temporal pattern that appears when the zero steady state loses stability, nonlinear terms of the original system must be taken into account, and a singular perturbation analysis around a bifurcation point must be pursued.

In the following we investigate the possible spatial and spatio-temporal patterns that can occur in the neuronal system with adaptation (1.3), as we vary the parameters α, g, τ, and θ.

Based on the hypotheses $F(0) = 0$, $F'(0) = 1$, the expansion of (1.3) in linear and higher order terms becomes

$$\begin{aligned} \frac{\partial u}{\partial t} &= -u + (\alpha J * u - gv) + \frac{F''(0)}{2}(\alpha J * u - gv)^2 + \frac{F'''(0)}{6}(\alpha J * u - gv)^3 + \cdots , \\ \frac{\partial v}{\partial t} &= (-v + u)/\tau , \end{aligned} \tag{2.1}$$

and then the linear operator is

$$L_0 U = \frac{\partial}{\partial t} U - \begin{pmatrix} -1 + \alpha J * (\cdot) & -g \\ 1/\tau & -1/\tau \end{pmatrix} U , \tag{2.2}$$

where $U = (u, v)^T$. We look for solutions of $L_0 U = \mathbf{0}$ that are bounded and have the form $\xi(t)\, e^{ikx}$ with $k \in \mathbb{R}$.

Let us assume first that $l = \infty$. Then, according to (1.4) and (1.5), equation (2.2) can be written as $\left[\frac{d\xi}{dt} - \hat{L}(k)\xi(t)\right] e^{ikx} = \mathbf{0}$, where

$$\hat{L}(k) = \begin{pmatrix} -1 + \alpha \hat{J}(k) & -g \\ 1/\tau & -1/\tau \end{pmatrix} . \tag{2.3}$$

Since we work on an infinite domain ($l = \infty$) and J is symmetric, this statement is true for all values of $k \in \mathbb{R}$. Moreover, we have $\hat{J}(-k) = \hat{J}(k)$.

The equation to be solved now is the ODE $\frac{d\xi}{dt} = \hat{L}(k)\xi$, which has two independent solutions $\xi_{1k}\, e^{\lambda_{1k} t}, \xi_{2k}\, e^{\lambda_{2k} t}$, where $\xi_{1,2\,k}$ are two-dimensional complex vectors. Therefore the eigenfunctions of L_0 have the form $\xi_{1,2\,k}\, e^{\lambda_{1,2\,k} t \pm ikx}$ and $\overline{\xi}_{1,2\,k}\, e^{\overline{\lambda}_{1,2\,k} t \mp ikx}$, where $\lambda_{1,2\,k}$ are the eigenvalues defined by

$$\lambda_{1,2\,k} = \frac{1}{2}\left[Tr(\hat{L}(k)) \pm \sqrt{Tr(\hat{L}(k))^2 - 4\det(\hat{L}(k))} \right] .$$

If $\det(\hat{L}(k)) > 0$ and $Tr(\hat{L}(k)) < 0$ for all k, i.e., $\alpha\, \hat{J}(k) < g + 1$ and $\alpha\, \hat{J}(k) < 1/\tau + 1$, then all eigenfunctions of L_0 lie on the stable manifold and decay exponentially in time to zero. The trivial solution is asymptotically stable.

Remark 4. The eigenvalues k represent a measure of the wave-like pattern that can occur in the system. That is why k are called *wavenumbers*, or *modes*, of the system, and $2\pi/k$ are called *wavelengths*.

We consider k_0 to be *the most unstable mode*, defined as

$$\hat{J}(k_0) = \max_{k\geq 0}\hat{J}(k) = \max_{k\geq 0}\left(\int_{-\infty}^{\infty} J(x)\, e^{ikx}\, dx\right), \tag{2.4}$$

and assume that

$$k_0 \neq 0 \quad \text{and} \quad \hat{J}(k_0) > 0\,, \tag{2.5}$$

$$\hat{J}(k_0) \neq \hat{J}(k)\ \forall k \neq \pm k_0\,. \tag{2.6}$$

This is true for functions J as in Figure 1.

There are only two ways the trivial solution can lose its stability: either when the determinant becomes zero or when the trace becomes zero. We notice that (2.4), with additional conditions (2.5), (2.6), implies that $Tr(\hat{L}(k)) < Tr(\hat{L}(k_0))$ and $\det(\hat{L}(k)) > \det(\hat{L}(k_0))$ for $k \neq \pm k_0$. Therefore k_0 is the first eigenvalue in which the system may lose its stability; that is, k_0 is the most unstable mode of system (1.3). For all $k \neq \pm k_0$ the eigenfunctions belong to the stable manifold. On the other hand, the eigenfunctions with $\pm k_0$ wavenumber may form a basis for the center manifold that becomes our point of interest.

The wavenumber k_0 determines then the mechanism that generates the emerged pattern. There are basically two possible cases. At $\alpha\,\hat{J}(k_0) = g+1$, $g < 1/\tau$, the determinant becomes zero and a spatial pattern (steady state (SS)) bifurcates. At $\alpha\,\hat{J}(k_0) = 1 + 1/\tau$, $g > 1/\tau$, the trace becomes zero and a spatio-temporal pattern (traveling wave (TW)/standing wave (SW)) bifurcates.

Let us assume now that l is finite. There is a considerable difference between working with finite domains as the interval $[-l, l]$ and periodic boundary conditions. The difference comes from the fact that in this case there is only a discrete set of possible wavenumbers. The wavenumbers k must satisfy the condition $k \in \left(\pm\frac{\pi}{l}\mathbb{N}\right)$ in order for the integral $\int_{-l}^{l} J(x-y)\, e^{ik(x-y)}\, dy$ to be independent of x and so equal to $\hat{J}(k) = \int_{-l}^{l} J(y)\, e^{iky}\, dy$.

This allows us to use the matrix $\hat{L}(k)$ from (2.3) and construct the eigenvalues and eigenfunctions of the linear operator, as in the case of infinite domain. Moreover, the discussion from the previous paragraph remains valid here with the observation that in the case of l finite we consider *only* those values of k belonging to the set $\left(\pm\frac{\pi}{l}\mathbb{N}\right)$.

The most unstable mode k_0 is then defined as

$$\hat{J}(k_0) = \hat{J}\left(\frac{\pi n_0}{l}\right) = \max_{k\in\frac{\pi\mathbb{N}}{l}}\left(\int_{-l}^{l} J(x)\, e^{ikx}\, dx\right) \tag{2.7}$$

and we assume again that $k_0 \neq 0$ and $\hat{J}(k_0) > 0$, $\hat{J}(k_0) \neq \hat{J}(k)\ \forall k = \pm\pi n/l$, $n \in \mathbb{N}$, such that $n \neq n_0$. This is true for functions J as in Figure 2.

Remark 5. In the following sections we analyze the case of spatial and spatio-temporal patterns that occur in the system when, at the most unstable mode $k_0 \neq 0$, either the trace $Tr(\hat{L}(k_0))$ becomes zero, the determinant $\det(\hat{L}(k_0))$ becomes zero, or *both* the trace $Tr(\hat{L}(k_0))$ and the determinant $\det(\hat{L}(k_0))$ become zero at the same value of α.

2.1. Two-population model with adaptation. Here we show that the linearization of the full two-population model does not differ substantially from our fast-inhibition simplification. Reconsider (1.2). Rescale time so that $\tau_E = 1$ and the linearized equations are

$$\begin{aligned}
\frac{du_E}{dt} &= -u_E + \alpha_E (J_{EE} * u_E - J_{EI} * u_I - gA)\,, \\
\tau_I \frac{du_I}{dt} &= -u_I + \alpha_I J_{IE} * u_E\,, \\
\tau_A \frac{dA}{dt} &= -A + u_E\,.
\end{aligned}$$

Here $\alpha_{E,I}$ are the derivatives of $F_{E,I}$ at the constant steady state. As above, the stability is determined by analyzing the eigenvalues of the matrix:

$$M(k) = \begin{pmatrix} -1 + \hat{J}_{EE}(k) & -\hat{J}_{EI}(k) & -g \\ \hat{J}_{IE}(k)/\tau_I & -1/\tau_I & 0 \\ 1/\tau_A & 0 & -1/\tau_A \end{pmatrix},$$

where we have absorbed the parameters $\alpha_{E,I}$ into the $\hat{J}$'s and the parameter g. The eigenvalues of this satisfy a cubic polynomial, $P(\lambda) = \lambda^3 + c_2\lambda^2 + c_1\lambda + c_0$. There is a zero eigenvalue when $c_0 = 0$ and there is an imaginary eigenvalue when $c_1 c_2 - c_0 = 0$. For example, the condition for a zero eigenvalue is

$$g = g_{crit} \equiv -1 + (\hat{J}_{EE}(k) - \hat{J}_{EI}(k)\hat{J}_{IE}(k))\,.$$

Note that if we define $\hat{J}(k) = \hat{J}_{EE}(k) - \hat{J}_{EI}(k)\hat{J}_{IE}(k)$, then this is the same condition we obtained in the simpler model. The double zero eigenvalue occurs when both $c_0 = 0$ and $c_1 = 0$, which yields the following condition on the time-constant of adaptation:

$$\tau_A = \frac{1 - \tau_I[\hat{J}_{IE}(k)\hat{J}_{EI}(k)]}{g_{crit}}\,,$$

which, if $\tau_I = 0$, is the same as our earlier condition. The condition for a Hopf bifurcation is more complicated but depends linearly on the degree of adaptation. For small τ_I, it reduces to the conditions above. Thus, for the simple zero and the Takens–Bogdanov bifurcation, the full three-variable model leads to nearly identical conditions on the parameters. The condition for a Hopf bifurcation is more complicated; the following must hold:

$$\begin{aligned}
(\hat{J}(k) - 1 - g)\tau_I\tau_A = {} & [\tau_A + \tau_I - \tau_A\tau_I(1 - \hat{J}_{EE}(k))] \\
& \times [\tau_A\hat{J}(k) - 1 - \tau_A + \tau_I(\hat{J}_{EE}(k) - 1 - g)]\,.
\end{aligned}$$

Note that, like the conditions for the Takens–Bogdanov bifurcation, the quantities depend only on $\hat{J}_{EE}$ and $\hat{J}$, the effective steady-state kernel. As $\tau_I \to 0$, this leads to $-1 + \hat{J}(k) = 1/\tau_A$, which was the condition for the reduced model. Since the linearized behavior of the full three-variable model is similar to that of the simplified system, we compute the normal form only for the simpler system. We expect that there will be little difference in the behavior of the full system.

3. Spatio-temporal patterns obtained by a loss of stability at a purely imaginary pair of eigenvalues. In the case of $Tr(\hat{L}(k_0)) = 0$ and $\det(\hat{L}(k_0)) > 0$, at the most unstable mode k_0 defined by (2.4), or (2.7), with conditions (2.5), (2.6), the eigenvalues of the associated ODE $\frac{d\xi}{dt} = \hat{L}(k_0)\xi$ are complex with zero real part. This happens when the parameters of system (1.3) satisfy

$$g > 1/\tau \quad \text{and} \quad \alpha^* = \frac{1 + 1/\tau}{\hat{J}(k_0)}\,. \tag{3.1}$$

Remark 6. In the following we fix the values of τ and g as above and take α as the bifurcation parameter. The bifurcation value around which we will consider the singular perturbation analysis is α^*. Therefore in the entire section 3, the operator L_0 defined by (2.2), and the matrix $\hat{L}(k)$ defined by (2.3) for all k, where it makes sense, will be evaluated at $\alpha = \alpha^*$.

The matrix $\hat{L}(k_0)$ has purely imaginary eigenvalues $\pm i\omega_0$ with corresponding eigenvectors Φ_0 and $\overline{\Phi}_0$ such that

$$\omega_0 = \frac{1}{\tau}\sqrt{g\tau - 1}\,, \tag{3.2}$$

$$\hat{L}(k_0)\Phi_0 = i\omega_0\Phi_0 \quad \text{with} \quad \Phi_0 = \left(\phi\,, \frac{\phi}{1 + i\sqrt{g\tau - 1}}\right)^T. \tag{3.3}$$

Based on the general theory [10], in the case of a pair of purely imaginary eigenvalues that arises at the most unstable mode k_0, the solution U of nonlinear system (1.3) can be approximated by

$$U(x,t) \approx 2\mathrm{Re}\left[z(t)\,\Phi_0\,e^{i(\omega_0 t + k_0 x)} + w(t)\,\Phi_0\,e^{i(\omega_0 t - k_0 x)}\right], \tag{3.4}$$

where z, w are time-dependent functions that satisfy the ODE system

$$\begin{cases} z' = z(a + bz\overline{z} + cw\overline{w})\,, \\ w' = w(a + bw\overline{w} + cz\overline{z}) \end{cases} \tag{3.5}$$

called *the normal form for the Turing–Hopf bifurcation* in the time-and-space-variable case, with $a = a_1 + ia_2$, $b = b_1 + ib_2$, and $c = c_1 + ic_2$ complex coefficients.

The importance of the normal form becomes apparent when we write it in polar coordinates. It provide us with essential information about the existence and stability of the (new) bifurcating solutions. We notice that actually only the signs and values of the real parts a_1, b_1, c_1 of the coefficients a, b, and c play a role in the matter.

In that sense let us define $z(t) = re^{i\theta_1}$ and $w(t) = Re^{i\theta_2}$. Then (3.5) is equivalent to the system $r' = r[a_1 + b_1r^2 + c_1R^2]$, $R' = R[a_1 + b_1R^2 + c_1r^2]$, $\theta_1' = a_2 + b_2r^2 + c_2R^2$,

$\theta_2' = a_2 + b_2 R^2 + c_2 r^2$, and the normal form is basically reduced to

$$r' = r\,[\,a_1 + b_1 r^2 + c_1 R^2\,]\,,\quad R' = R\,[\,a_1 + b_1 R^2 + c_1 r^2\,]\,. \tag{3.6}$$

There are two distinct qualitative pictures of small amplitude bifurcating patterns in a system with normal form (3.5), and so (3.6), as long as $-a_1/b_1 > 0$ and $-a_1/(b_1+c_1) > 0$ (see Ermentrout [10]).

One corresponds to the solution $\tilde{r} = 0$, $\tilde{R} = \sqrt{-a_1/b_1}$ of (3.6) (or $\tilde{R} = 0$, $\tilde{r} = \sqrt{-a_1/b_1}$) and it represents a traveling periodic wave train with velocity $c = \pm\omega_0/k_0$ ("*traveling wave*"). This can be understood easily by using formula (3.4): Up to a translation in time, an approximation of the solution $U(x,t)$ is then

$$2\sqrt{-a_1/b_1}\mathrm{Re}\left[\Phi_0\, e^{i(\omega_0 t \pm k_0 x)}\right] = 2\sqrt{-a_1/b_1}\mathrm{Re}\left[\Phi_0\, e^{\mp i k_0 (ct - x)}\right]. \tag{3.7}$$

Therefore the pattern will change with time and position in space, according to the traveling wave coordinate $\xi = ct - x$ (for an example, see Figure 6).

The other case corresponds to the solution $\tilde{r} = \tilde{R} = \sqrt{-a_1/(b_1+c_1)}$, and it represents a standing oscillation, periodic in space with spatial frequency k_0, and periodic in time with temporal frequency ω_0 ("*standing wave*"). The approximating solution of $U(x,t)$ (up to a translation in time) is now

$$4\sqrt{-a_1/(b_1+c_1)}\mathrm{Re}\left[\Phi_0\, e^{i\omega_0 t}\right]\cos(k_0 x)\,, \tag{3.8}$$

and the pattern consists of oscillations with respect to the position x in space for any fixed time t or in oscillations with respect to time at any fixed position x (for an example, see Figure 5).

They cannot be simultaneously stable; therefore, physically, only one of these patterns is selected [10], [9], [12]. The traveling wave solution TW has the corresponding eigenvalues $\lambda_1 = -2a_1$, $\lambda_2 = -\frac{a_1(c_1-b_1)}{b_1}$ with eigenvectors $(1,0)^T$, $(0,1)^T$. Therefore *the traveling wave exists and it is stable* if and only if $a_1 > 0$, $b_1 < 0$, and $c_1 - b_1 < 0$. The standing wave solution SW has the corresponding eigenvalues $\lambda_1 = -2a_1$, $\lambda_2 = -\frac{2a_1(b_1-c_1)}{b_1+c_1}$ with eigenvectors $(1,1)^T$ and $(1,-1)^T$. Therefore *the standing wave exists and it is stable* if and only if $a_1 > 0$, $b_1 + c_1 < 0$, and $c_1 - b_1 > 0$.

We summarize the above observations into the following two theorems.

Theorem 3.1 (existence and stability of traveling and standing waves through a Turing–Hopf bifurcation. See [10]). *Let us consider the normal form* (3.5) *for the Turing–Hopf bifurcation of a nonlinear (time- and one-dimensional-space-dependent) system, and define* $a_1 = Re(a)$, $b_1 = Re(b)$, *and* $c_1 = Re(c)$.

(i) *System* (3.5) *has a TW solution if and only if* $(-a_1)/b_1 > 0$. *Moreover, the TW exists and it is stable if and only if*

$$a_1 > 0\,,\ b_1 < 0\,,\ c_1 - b_1 < 0\,.$$

(ii) *System* (3.5) *has an SW solution if and only if* $(-a_1)/(b_1+c_1) > 0$*. Moreover, the SW exists and it is stable if and only if*

$$a_1 > 0\,,\ b_1 + c_1 < 0\,,\ c_1 - b_1 > 0\,.$$

Theorem 3.2 (see [10]). *Let us assume that a nonlinear (time- and one-dimensional-space-dependent) system passes at the most unstable mode* k_0 *through a Turing–Hopf bifurcation, with eigenvalues* $\pm i\omega_0$ *and eigenvectors* $\Phi_0, \overline{\Phi}_0$*. The associated normal form is* (3.5)*. Then*

(i) *the linear approximation of the TW solution is* (3.7)*. The velocity of the TW is* $\pm\omega_0/k_0$*.*

(ii) *The linear approximation of the SW solution is* (3.8)*.*

Since the goal of our study is to investigate the existence of stable TW and/or SW patterns in the neural network (1.3), we have to construct the normal form for the Hopf bifurcation case. More precisely, according to the general theory summarized above, we have to determine the coefficients a, b, and c in (3.5) and then determine their real parts.

3.1. Traveling wave and standing wave patterns in the neural system. The construction of the normal form uses a singular perturbation approach with a proper scaling of the variables, parameters, and time with respect to ϵ, the small pertubation quantity. The Fredholm alternative method is then used to identify solutions for the functional equations obtained from the ϵ-power series expansion.

We include in this section only the main results obtained as a consequence of the construction of the Turing–Hopf bifurcation normal form. A summary of the basic steps of perturbation calculations is then included in Appendix A.

Therefore we are able to compute the expression of the coefficients of the normal form as a dependence on the original parameters of the nonlinear system (1.3). That allows us to identify the regions in the parameter space where stable traveling wave or stable standing wave patterns occur in system (1.3).

We should mention that the numerical simulations of the full nonlinear model around a Hopf bifurcation point indeed showed the presence of stable traveling waves. Nevertheless it was a difficult task to find in simulations the region in the parameter space corresponding to *stable standing waves*. Therefore in this case the analysis is a must. The computed coefficients of the normal form help us to prove the existence of stable standing waves as solutions of the neural system (1.3), and to visualize them numerically.

We state below the results obtained from the normal form calculation.

Theorem 3.3. *If* $g > 1/\tau$*, in the neighborhood of the bifurcation value* $\alpha^* = \frac{1+1/\tau}{\hat{J}(k_0)}$*, system* (1.3) *has the normal form* (3.5) *with*

$$a_1 = Re(a) = \frac{1}{2}\left[\alpha\,\hat{J}(k_0) - \left(1 + \frac{1}{\tau}\right)\right],$$

and $b_1 = Re(b)$*,* $c_1 = Re(c)$ *satisfying the equations*

$$b_1 = \frac{\tau+1}{4\tau}\,|A|^2\left[F'''(0) + F''(0)^2\cdot\left(-3 + \frac{2}{g+1-\frac{(1+1/\tau)\hat{J}(0)}{\hat{J}(k_0)}} + \frac{M_B}{N_B}\right)\right], \tag{3.9}$$

$$
\begin{aligned}
c_1 + b_1 = \frac{\tau+1}{4\tau}\,|A|^2 \cdot \Bigg[& 3\left[F'''(0) - 3F''(0)^2\right] \\
& + F''(0)^2 \cdot \left(\frac{2}{g+1-\frac{(1+1/\tau)\hat{J}(2k_0)}{\hat{J}(k_0)}} + \frac{4}{g+1-\frac{(1+1/\tau)\hat{J}(0)}{\hat{J}(k_0)}} + 2\frac{M_C}{N_C} + \frac{M_B}{N_B} \right) \Bigg],
\end{aligned} \tag{3.10}
$$

$$
\begin{aligned}
c_1 - b_1 = \frac{\tau+1}{4\tau}\,|A|^2 \cdot \Bigg[& \left[F'''(0) - 3F''(0)^2\right] \\
& + F''(0)^2 \cdot \left(\frac{2}{g+1-\frac{(1+1/\tau)\hat{J}(2k_0)}{\hat{J}(k_0)}} + 2\frac{M_C}{N_C} - \frac{M_B}{N_B} \right) \Bigg].
\end{aligned} \tag{3.11}
$$

Here we have $M_B = M\big(\frac{\hat{J}(2k_0)}{\hat{J}(k_0)}\big)$, $M_C = M\big(\frac{\hat{J}(0)}{\hat{J}(k_0)}\big)$, $N_B = N\big(\frac{\hat{J}(2k_0)}{\hat{J}(k_0)}\big)$, $N_C = N\big(\frac{\hat{J}(0)}{\hat{J}(k_0)}\big)$, *where* M *and* N *are functions defined as*

$$M(X) = (4g\tau - 3)[2g\tau - (\tau+1)(\tau+2)]X + 4(g\tau-1)(\tau+1)^2 + (3g\tau - 4 - \tau)^2 + g\tau(g\tau + \tau - 2)$$

and

$$N(X) = (4g\tau-3)(\tau+1)^2X^2 + 2\tau(\tau+1)(3-g-4g\tau)X + [4(g\tau-1)(\tau+1)^2 + (3g\tau-4-\tau)^2]\,.$$

Based on formulas (3.9), (3.10), (3.11) from Theorem 3.3, we obtain the first important result regarding the type of patterns that can be selected by neural system (1.3).

Theorem 3.4. *Let us assume that the most unstable mode* k_0 *of system* (1.3) *satisfies conditions* (2.5), (2.6), *and at* k_0 *a pair of purely imaginary eigenvalues appears.*

If the firing-rate function F *is such that* $F(0) = 0$, $F'(0) > 0$, $F''(0) = 0$, *and* $F'''(0) < 0$, *then system* (1.3) *has a TW and an SW solution for* $\alpha > \alpha^*$, α *close to* α^*. *The SW solution is unstable. The TW solution is stable.*

Proof. First we notice that there exist sigmoid functions F that satisfy the theorem hypotheses. For example, if $\theta = 0$, we have from (1.8) $F(u) = \frac{2}{r}\tanh\left(\frac{ru}{2}\right)$, and thus $F(0) = 0$, $F'(0) = 1$, $F''(0) = 0$, $F'''(0) = -\frac{r^2}{2} < 0$.

In this case $b_1 = \frac{\tau+1}{4\tau}|A|^2F'''(0) < 0$, $c_1 = 2b_1$, and therefore $c_1 + b_1 < 0$ and $c_1 - b_1 < 0$. By Theorem 3.1, both SWs and TWs bifurcate from the trivial solution at $\alpha = \alpha^*$, but only the TWs are stable. ■

Remark 7. Condition $F''(0) = 0$ on the firing-rate function is quite restrictive. We are interested in seeing what happens in the general case of $F''(0) \neq 0$ (then the firing-rate function F as in (1.8) has the second and third derivatives $F''(0) = r\,\frac{1-e^{-r\theta}}{1+e^{-r\theta}}$, $F'''(0) = \frac{r^2\left(e^{-2r\theta}-4e^{-r\theta}+1\right)}{(1+e^{-r\theta})^2}$). In that sense, the coefficients of the normal form (3.5) computed in the previous section provide us with some useful information. They have indeed a complicated expression that does not allow us to give a general, theoretical prediction. Nevertheless we can use (3.9), (3.10), (3.11) in MATLAB, for example, to search for possible parameter values of g, τ, r, θ, plus coupling J, such that the stable pattern *selected* in the bifurcation at α^* is an SW. The importance of the construction of the normal form (3.5) becomes clear since this allows us to show that *both* TW *and* SW patterns can be found in the neural system (1.3).

Theorem 3.5. *Let us assume that the hypotheses in Theorem 3.3 are true.*

(i) *If $b_1 < 0$ and $c_1 - b_1 < 0$, then for $\alpha > \alpha^*$, sufficiently close to α^*, system (1.3) has a TW solution that is stable. The velocity of the TW is approximately $\left(\pm\frac{\sqrt{g\tau-1}}{\tau k_0}\right)$, and the solution can be approximated by*

$$U(x,t) \approx \sqrt{\frac{2\hat{J}(k_0)(\alpha-\alpha^*)}{(-b_1)}}\, Re\left[\Phi_0 e^{i(\omega_0 t \pm k_0 x)}\right]$$

with Φ_0 and ω_0 defined by (3.2) and (3.3).

(ii) *If $c_1 + b_1 < 0$ and $c_1 - b_1 > 0$, then for $\alpha > \alpha^*$, sufficiently close to α^*, system (1.3) has an SW solution that is stable. The solution can be approximated by*

$$U(x,t) \approx 2\cos(k_0 x)\sqrt{\frac{2\hat{J}(k_0)(\alpha-\alpha^*)}{(-b_1-c_1)}}\, Re\left[\Phi_0 e^{i\omega_0 t}\right].$$

Proof. The proof follows immediately from Theorems 3.2 and 3.3. ■

Example. If we consider an infinite domain ($l = \infty$), the synaptic coupling J is defined by (1.6) with a graph, as in Figure 1. For example, for $A = 5$, $B = 4$, $a = 1$, $b = 0.3$ we have $k_0 = 1.2967$, $\hat{J}(0) = 1$, $\hat{J}(k_0) = 2.2988$, $\hat{J}(2k_0) = 0.9158$, and at $\tau = 4$ we obtain $\alpha^* = 0.5438$. We choose the function F as in (1.8) with $r = 3$ and $\theta = 0.3$. The theory predicts that there exist values of g such that the stable pattern in neural network (1.3) that occurs through the Hopf bifurcation is the TW, and there exist values of g such that the SW pattern is stable. For example, at $g = 0.34$ both TW and SW bifurcate, but only SW is stable ($b_1 = -0.0651$, $c_1 + b_1 = -0.0955$, $c_1 - b_1 = 0.0347$). On the other hand, at $g = 0.35$ both TW and SW bifurcate, but only TW is stable ($b_1 = -0.1283$, $c_1 + b_1 = -0.2873$, $c_1 - b_1 = -0.0306$).

Remark 8. For the infinite domain problem, the normal form is often formally supplemented with long wave modulation equations. This is a consequence of the fact that, when the bifurcation parameter crosses criticality, an entire *band* of values of the wavenumber k becomes unstable. Formally, a spatial variable $X = \epsilon x$ is introduced and the normal form variables w, z are functions of both the slow time and the slow space scale. The normal form becomes

$$\begin{aligned} z' &= z(a + bz\bar{z} + cw\bar{w}) + dz_{XX} + ew_{XX}\,, \\ w' &= w(a + bw\bar{w} + cz\bar{z}) + dw_{XX} + ez_{XX}\,, \end{aligned}$$

where d, e are complex parameters which depend on the second derivative of the function $\hat{J}(k)$ at the critical wavenumber [17]. These modulation equations are not relevant in the finite size domain since the wavenumbers k take only discrete values.

3.2. Numerical results.

A good agreement is obtained between the theoretical prediction (based on the normal form construction) and the numerical simulation of the full nonlinear system (1.3).

For numerical simulations we need to consider a finite domain together with periodic boundary conditions. The synaptic coupling J is defined by (1.7) with a graph as in Figure 2, and there is only a discrete set of wavenumbers. We choose the gain function F as in (1.8) with $r = 3$ and $\theta = 0.3$, or $\theta = 0$, and $l = \pi$ and $a = -0.2$, $b = 2.5$, $c = 2$ in J. Then $k_0 = 1$, $\hat{J}(0) = -0.2$, $\hat{J}(k_0) = 1.25$, $\hat{J}(2k_0) = 1$, and at $\tau = 4$ we obtain $\alpha^* = 1$. The simulations for system (1.3) were run in XPPAUT [7] on a network of 100 neurons, with the method of integration Runge–Kutta RK4 and step size $dt = 0.25$.

At $\theta = 0.3$, the theory predicts that, for example, at $g = 0.45$, both TW and SW bifurcate, but only SW is stable ($b_1 = -3.4412$, $c_1+b_1 = -5.1928$, $c_1-b_1 = 1.6895$). At $g = 0.7$, both TW and SW bifurcate, but only TW is stable ($b_1 = -3.1939$, $c_1+b_1 = -7.7540$, $c_1-b_1 = -1.3661$).

At $\theta = 0$, for any $g > 1/\tau = 0.25$, we can obtain both SW and TW solutions, but the stable pattern is always TW.

These results are confirmed by the numerical simulations of the full model (1.3).

Remark 9. In the figures below we represent the space x on the horizontal axis, the time t on the vertical axis, and the value of the variable $u(x, t)$ by the level of color. The upper left corner corresponds to the minimum value of x that increases to the right. The value of time increases in the up-to-down direction. In general, the time is represented after $t = 3000$ transients. We choose two different sets of initial conditions to illustrate the possible behaviors in system (1.3).

Let us consider first the case $\theta = 0.3$.

For different values of the parameter g, e.g., $g = 0.45$ and $g = 0.7$, before the bifurcation point, at $\alpha = 0.99$, by choosing random initial conditions around the origin, the solution decays in time to zero. After the bifurcation point, at $\alpha = 1.01$, both TW and SW patterns can be obtained, depending on the choice of the initial conditions. In order to test which pattern is stable, we have also run the simulations of system (1.3) in the presence of white noise added to the first equation and scaled by a factor of 0.001 (see Figures 4(b)–7(b)).

As a consequence we notice that at $g = 0.45$, the stable pattern is SW (see Figures 4 and 5), and at $g = 0.7$ the stable pattern is TW (see Figures 6 and 7). This means that for both sets of initial conditions—that in the absence of noise might produce different patterns—we obtain *the same pattern in the presence of noise*, that is, SW at $g = 0.45$, respectively, TW at $g = 0.7$.

At $\theta = 0$ the stable pattern obtained as a result of added noise is always TW. We present the numerical results in Figures 8 and 9 for $g = 0.45$, and in Figures 10 and 11, respectively, for $g = 0.7$.

3.3. Summary. We have shown that with strong adaptation, $g\tau > 1$, spatio-temporal patterns bifurcate from the uniform resting state. Depending on the position of the rest state relative to the inflection point of the sigmoid nonlinearity, we can get either TWs or standing oscillations. The latter were not found in the model of Hansel and Sompolinsky (or at least they were not discussed). Ermentrout [10] found both TWs and SWs in the Wilson–Cowan equations but did not compute the normal forms needed to specify which were stable.

The biological interpretation of the TWs is fairly clear. There are many examples of such TWs in experimental preparations. These are reviewed and their possible functional role is discussed in [13]. SWs are more difficult to find in experimental preparations. However, suppose that we view the model in the manner of Hansel and Sompolinsky, where the spatial

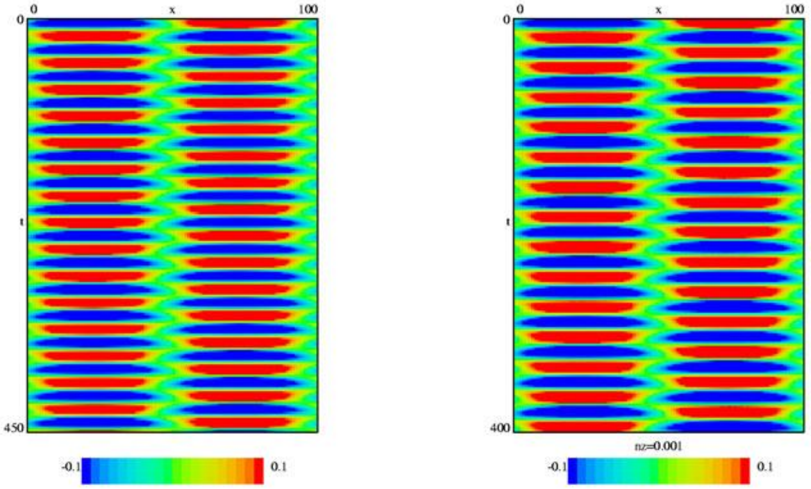


Figure 4. (a) *SW is the pattern obtained at $g = 0.45$, $\theta = 0.3$, and set 1 of initial conditions.* (b) *In the presence of noise, SW is preserved.*

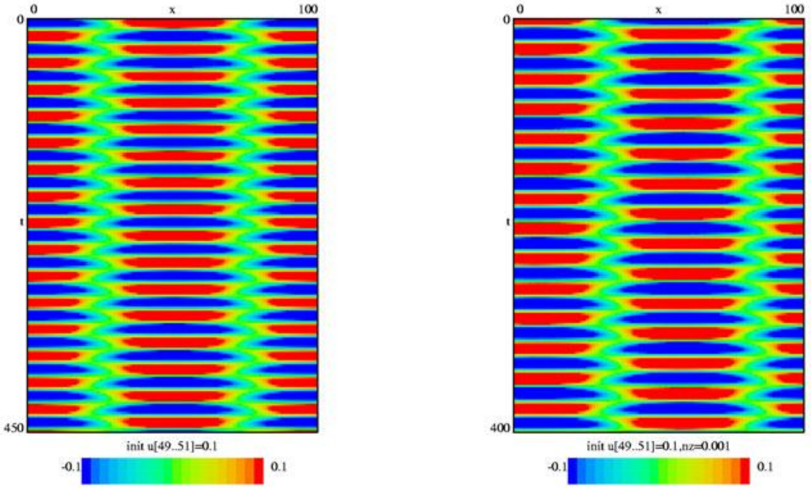


Figure 5. (a) *SW is the pattern obtained at $g = 0.45$, $\theta = 0.3$, and set 2 of initial conditions.* (b) *It is stable since in the presence of noise, SW is preserved.*

variable x represents the orientation tuning of a cell. Suppose that the critical wavenumber is 1 so that the SW solution has the form $\cos\omega t \cos 2\pi x/L$. This represents two groups of cells whose preferred orientations are mutually orthogonal; when one group is at its maximum, the other is at its minimum. Solutions of this form have been posited as neural analogues of

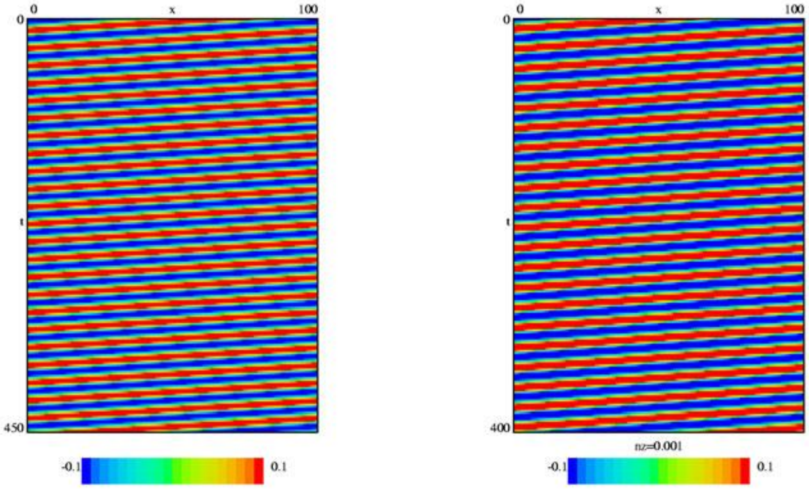


Figure 6. (a) *TW is the pattern obtained at $g = 0.7$, $\theta = 0.3$, and set* 1 *of initial conditions.* (b) *In the presence of noise, TW is preserved.*

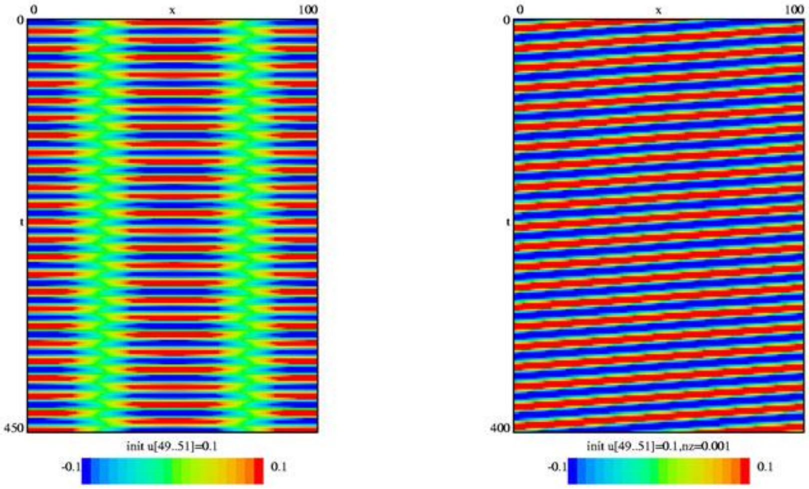


Figure 7. (a) *SW is the pattern obtained at $g = 0.7$, $\theta = 0.3$, and set* 2 *of initial conditions.* (b) *It is unstable since in the presence of noise, SW is replaced by TW.*

perceptual reversals [23] and binocular rivalry [18].

4. Spatial patterns obtained by a loss of stability at zero eigenvalue. The case when the determinant vanishes first and the trace is still negative is analyzed in this section. It

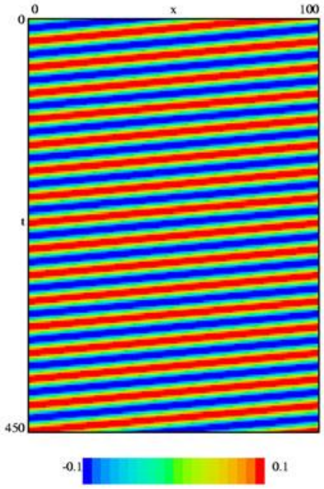


Figure 8. *TW is the pattern obtained at $g = 0.45$, $\theta = 0$, and set 1 of initial conditions.*

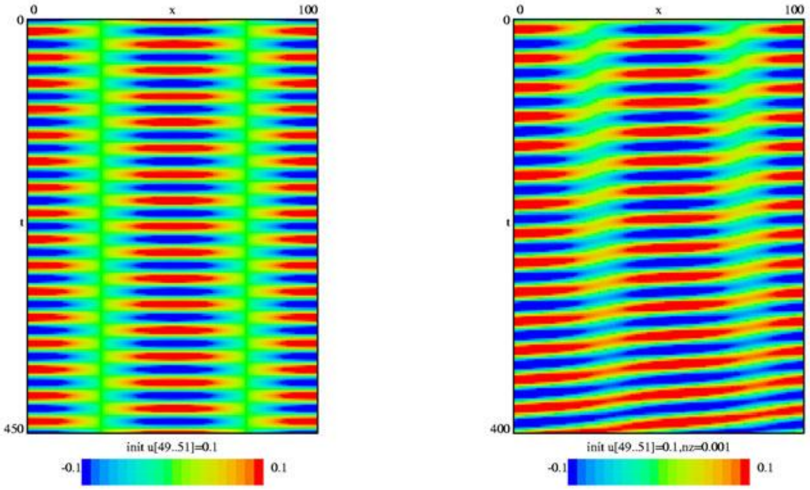


Figure 9. (a) *SW is the pattern obtained at $g = 0.45$, $\theta = 0$, and set 2 of initial conditions.* (b) *It is unstable since in the presence of noise, SW is replaced by TW.*

corresponds to a choice of parameters of system (1.3) satisfying the following conditions:

$$g < 1/\tau \quad \text{and} \quad \alpha^* = \frac{g+1}{\hat{J}(k_0)}\,. \tag{4.1}$$

At the most unstable mode k_0 we have $\det(\hat{L}(k_0)) = 0$ and $Tr(\hat{L}(k_0)) < 0$; therefore the

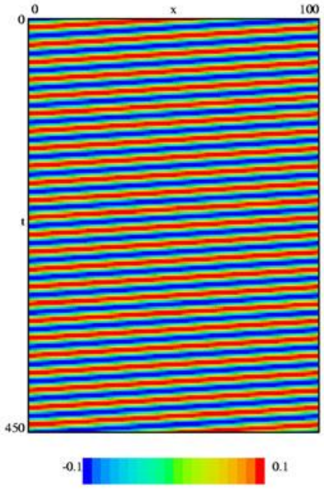


Figure 10. *TW is the pattern obtained at $g = 0.7$, $\theta = 0$, and set 1 of initial conditions.*

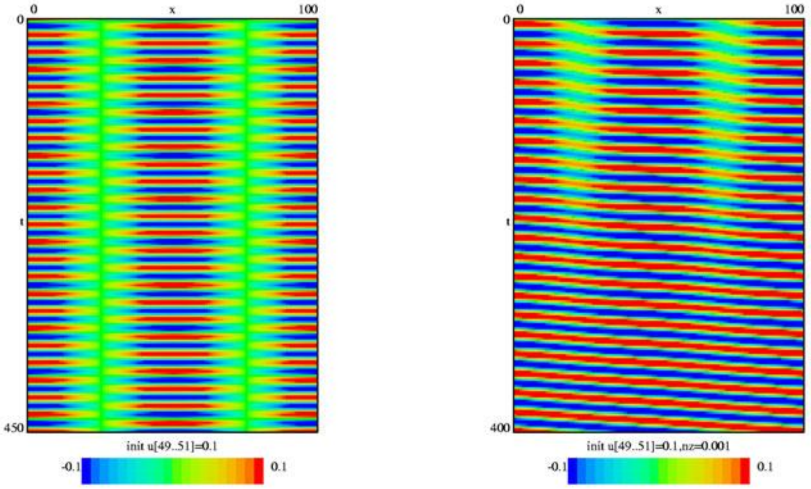


Figure 11. (a) *SW is the pattern obtained at $g = 0.7$, $\theta = 0$, and set 2 of initial conditions.* (b) *It is unstable since in the presence of noise, SW is replaced by TW.*

associated ODE system $\frac{d\xi}{dt} = \hat{L}(k_0)\xi$ has a simple zero eigenvalue with eigenvector Φ_0 defined by $\hat{L}(k_0)\Phi_0 = \mathbf{0}$, that is, for example, $\Phi_0 = (1,1)^T$. Let us consider also the eigenvector Ψ_0 of the adjoint equation $\hat{L}(k_0)^T\Psi_0 = \mathbf{0}$ such that $\Phi_0 \cdot \Psi_0 = 1$. Then we have $\Psi_0 = \frac{1}{1-g\tau}(1, -g\tau)^T$.

The pattern that results through the bifurcation at a simple zero eigenvalue oscillates with respect to space position due to the unstable mode k_0, but it is independent of time (the details of the construction of the normal form are given in Appendix A). We call this pattern *steady state* or *stationary pattern.*

Theorem 4.1. *If $g < 1/\tau$, in the neighborhood of the bifurcation value $\alpha^* = \frac{g+1}{\hat{J}(k_0)}$, system (1.3) has the normal form*

$$z' = \eta_1 z + \Lambda |z|^2 z \tag{4.2}$$

with $\eta_1 = \frac{\alpha\,\hat{J}(k_0)-(g+1)}{1-g\tau}$ and coefficient

$$\begin{aligned}\Lambda = {} & \frac{1}{2(1-g\tau)}\left[F'''(0) - 3F''(0)^2\right]\\ & + \frac{F''(0)^2}{(1-g\tau)(g+1)} \cdot \left[\frac{\hat{J}(k_0)}{\hat{J}(k_0)-\hat{J}(0)} + \frac{\hat{J}(k_0)}{2[\hat{J}(k_0)-\hat{J}(2k_0)]}\right].\end{aligned} \tag{4.3}$$

Theorem 4.2. *In the hypotheses of Theorem 4.1, the SS solution that occurs about the bifurcation point $\alpha^* = \frac{g+1}{\hat{J}(k_0)}$ has the following first order approximation:*

$$SS\ :\ u(x,t) = v(x,t) \approx 2\,\cos(k_0 x)\sqrt{\frac{\alpha\,\hat{J}(k_0) - (g+1)}{(1-g\tau)(-\Lambda)}}\,. \tag{4.4}$$

SS is stable if and only if $\Lambda < 0$ and $\alpha > \alpha^*$.

Proof. In polar coordinates $z = r\,e^{i\theta_1}$, the normal form equivalent to (4.2) is $r' = r(\eta_1 + \Lambda r^2)$, $\theta_1' = 0$. A nonzero solution exists only in the case $\eta_1\Lambda < 0$ and it is $\tilde{r} = \sqrt{\frac{\eta_1}{-\Lambda}}$ and θ_1 a constant.

The first order approximation of the nontrivial solution $U(x,t)$ is then

$$U(x,t) \approx z(t)\Phi_0 e^{ik_0x} + \overline{z}(t)\Phi_0 e^{-ik_0x} = 2\Phi_0 \mathrm{Re}\left[z(t)\,e^{ik_0x}\right] = 2\sqrt{\frac{\eta_1}{-\Lambda}}\Phi_0\cos(k_0x+\theta_1)\,.$$

Since $\Phi_0 = (1,1)^T$, we obtain exactly (4.4) up to a translation in space.

The stability condition comes from $h'(\tilde{r}) = -2\eta_1 < 0$, where $h(r) = r(\eta_1 + \Lambda r^2)$, together with the existence condition $\eta_1\Lambda < 0$. ■

We note that if the parameter θ in the firing-rate function F from (1.8) is $\theta = 0$, we obtain $F''(0) = 0$, $F'''(0) < 0$. Then $\Lambda < 0$ and the pitchfork bifurcation is supercritical: a (stable) stationary pattern occurs for $\alpha > \frac{g+1}{\hat{J}(k_0)}$.

5. Spatio-temporal patterns obtained by a loss of stability at a double-zero eigenvalue.

In the previous sections, we analyzed bifurcation to TWs and SWs when $g\tau > 1$ (strong adaptation) and to stationary patterns when $g\tau < 1$ (weak adaptation). The former (resp., latter) patterns occur when the trace becomes positive at a lower (resp., higher) value of α than at which the determinant becomes negative. We note that, in both cases, the normal form is no longer defined in the limit as $g\tau \to 1$; for the Hopf case, $\omega_0 \to 0$; and in the simple zero eigenvalue case, the coefficients of the normal form become unbounded.

An obvious question is how these possible patterns in system (1.3) interact, that is, how the system's behavior changes from TW or SW to a stationary pattern or vice versa. The transition between spatio-temporal and only spatial patterns can be analyzed by a study of the case when the trace and the determinant of the linearized system vanish simultaneously. Then at the most unstable mode we obtain a double-zero eigenvalue. The double-zero eigenvalue case is approached from two different directions: one in which we already have a zero eigenvalue and now obtain another (that is, coming from the domain of spatial/stationary patterns), and another in which we have a pair of purely imaginary eigenvalues $\pm i\omega_0$ that collide (that is, coming from the domain of spatio-temporal patterns).

As a result of the above remarks, the aim of the present section is to study behavior in system (1.3) at the transition between stationary states and TWs/SWs. Therefore we assume that at the most unstable mode k_0 we have $Tr(\hat{L}(k_0)) = \det(\hat{L}(k_0)) = 0$. This is true when the parameters satisfy the conditions

$$g^* = 1/\tau \quad \text{and} \quad \alpha^* = \frac{1 + 1/\tau}{\hat{J}(k_0)}\,. \tag{5.1}$$

Remark 10. In the following we fix the value of τ and take α and g as bifurcation parameters. The bifurcation values around which we will consider the singular perturbation analysis are α^* and g^*. Therefore in the entire section 5, the operator L_0 defined by (2.2), and the matrix $\hat{L}(k)$ defined by (2.3) for all k, where it makes sense, will be evaluated at $\alpha = \alpha^*$ and $g = g^*$.

At $\pm k_0$ the associated ODE $\frac{d\xi}{dt} = \hat{L}(k_0)\xi$ has a double-zero eigenvalue. For all other values $k \neq \pm k_0$ we have $Tr(\hat{L}(k)) < 0$ and $\det(\hat{L}(k)) > 0$, and the corresponding eigenvalues have negative real part.

Let us construct the (generalized) eigenvectors of $\hat{L}(k_0)$ and $\hat{L}(k_0)^T$ as follows:

$$\Phi_0 = \frac{1}{\sqrt{\tau}}\begin{pmatrix}1\\1\end{pmatrix},\ \Psi_1 = \frac{1}{\sqrt{\tau}}\begin{pmatrix}1\\-1\end{pmatrix},\ \Phi_1 = \sqrt{\tau}\begin{pmatrix}1\\0\end{pmatrix},\ \Psi_0 = \sqrt{\tau}\begin{pmatrix}0\\1\end{pmatrix}, \tag{5.2}$$

according to the conditions

$$\begin{cases}\hat{L}(k_0)\Phi_0 = \mathbf{0}\,,\ \hat{L}(k_0)\Phi_1 = \Phi_0\,,\ \hat{L}(k_0)^T\Psi_1 = \mathbf{0}\,,\ \hat{L}(k_0)^T\Psi_0 = \Psi_1\,,\\ \Phi_0\cdot\Psi_0 = \Phi_1\cdot\Psi_1 = 1\,,\ \Phi_0\cdot\Psi_1 = \Phi_1\cdot\Psi_0 = 0\,.\end{cases} \tag{5.3}$$

Around the double-zero bifurcation point, the first order approximation of the solution of nonlinear system (1.3) is given by its projection on the generalized eigenspace. That means U can be approximated by

$$U(x,t) \approx 2\mathrm{Re}\left[z(t)\,\Phi_0\,e^{ik_0x} + w(t)\,\Phi_1\,e^{ik_0x}\right], \tag{5.4}$$

where z, w are time-dependent functions that satisfy the ODE system with real coefficients

$$(5.5)\qquad \begin{cases} z' = w\,, \\ w' = \zeta_1 z + \zeta_2 w + A|z|^2 z + Cz[\,\overline{z}\,w + z\,\overline{w}\,] + D|z|^2 w\,, \end{cases}$$

called *the normal form for the double-zero (Takens–Bogdanov) bifurcation with* $O(2)$*-symmetry* [6].

Indeed the linear part of system (5.5) has two zero eigenvalues when the parameters $\zeta_1 = 0$ and $\zeta_2 = 0$. That is why we call this type of bifurcation "double-zero" or "Takens–Bogdanov." The additional name of "$O(2)$-symmetry" comes from the fact that system (5.5) exhibits symmetry under both rotations and reflections. This means that the vector field $G(z, w)$ (the right-hand side of (5.5)) commutes with the rotation $z \mapsto e^{i\theta}z$ for any angle $\theta \in \mathbb{R}$, i.e., we have $G(e^{i\theta}z, e^{i\theta}w) = e^{i\theta}G(z, w)$, and it also commutes with reflection $z \mapsto \overline{z}$, i.e., $G(\overline{z}, \overline{w}) = \overline{G(z, w)}$. Therefore the system shows no directional preference and we say that it is *isotropic*. The technical terminology is that the vector field G, and therefore system (5.5), is covariant (or equivariant) with respect to the group $O(2)$ of rotations and reflections.

Recall that for (1.3) we seek spatially periodic solutions of the form $\psi_k\, e^{ikx} + cc$, where cc stands for the complex conjugate of the previous term.

Since $\left[\psi_k\, e^{ik(x+d)} + cc = e^{ikd}\psi_k\, e^{ikx} + cc\right]$, a translation in space $[x \mapsto x+d]$ will be associated with a rotation of the time-dependent vector ψ_k. Furthermore, a reflection in space $[x \mapsto (-x)]$ is associated with a reflection of the vector ψ_k since $\left[\psi_k\, e^{ik(-x)} + cc = \overline{\psi}_k\, e^{ikx} + cc\right]$. We note that the vector field of original system (1.3) satisfies properties $G(\,u(x+d, t), v(x+d, t)\,) = G(u, v)(x + d, t)$ and $G(\,u(-x, t), v(-x, t)\,) = G(u, v)(-x, t)$, so we say that system (1.3) is isotropic.

System (1.3) has at least one solution that preserves the symmetry with respect to both rotations, $u(x + d, t) = e^{ikd}u(x, t)$, $v(x + d, t) = e^{ikd}v(x, t)$, and reflection $u(-x, t) = \overline{u(x, t)}$, $v(-x, t) = \overline{v(x, t)}$, and this is the trivial solution $u(x, t) = v(x, t) \equiv 0$. For different values of parameters, other solutions may exist which do not necessarily preserve the symmetry. We say that the symmetry in the system is broken and call the phenomenon that leads to this situation *symmetry breaking bifurcation*. The above-mentioned correspondence between (1.3) and (5.5), together with formula (5.4), allows us to work with system (5.5) and detect the solutions that break its symmetry, rather than working with (1.3).

Dangelmayr and Knobloch present in [6] a detailed analysis of the existence and stability properties for five types of possible solutions of system (5.5). These are the trivial solution/T, steady state/SS, traveling wave/TW, standing wave/SW, and modulated wave/MW. Depending on the sign of the coefficient A, and then the signs of D and $M = 2C + D$, together with some nondegeneracy conditions based on the value of the ratio D/M, different regions in the parameter plane (ζ_1, ζ_2) were identified, and the corresponding bifurcation diagrams were drawn. That is, as a dependence on the values of parameters, all possible qualitatively different behaviors in the system are described.

As an example, TWs break the symmetry with respect to reflection and keep the symmetry to rotations. On the other hand, SWs break the symmetry with respect to rotations but keep the symmetry to reflection (see below).

We summarize in the following the basic ideas followed by Dangelmayr and Knobloch [6] in their analysis. Moreover, in a similar approach to section 3 we give a geometric interpretation of the solutions SS, TW, SW, and MW.

First we write z and w in polar coordinates and transform system (5.5) accordingly. Since $w = z'$ we need only the polar representation of z, say $z(t) = r\,e^{i\phi}$. Then by the separation of the real and imaginary parts, system (5.5) is equivalent to

$$\begin{cases} r'' - r(\phi')^2 - r(\zeta_1 + Ar^2) - r'\,(\zeta_2 + Mr^2) = 0\,, \\ r\,\phi'' + 2r'\phi' - r\,\phi'\,(\zeta_2 + Dr^2) = 0\,. \end{cases} \tag{5.6}$$

The trivial solution T corresponds to the solution $r = 0$ and exists for all parameter values. Therefore $z(t) = w(t) = 0$ and from (5.4) we have $U(x,t) \equiv 0$. The solution is independent of time and position in space.

The linearization of (5.6) around $r = 0$, $\phi' = 0$, e.g., take $r = 0 + \xi$, $\phi' = 0 + \eta$ with ξ, η small, is the equation $\xi'' - \zeta_2\,\xi' - \zeta_1\,\xi = 0$. The eigenvalues have negative real part if and only if $\zeta_1 < 0$ and $\zeta_2 < 0$. The stability of the trivial solution T is lost at $\zeta_1 = 0$ through a zero eigenvalue when other constant solutions r_0 appear (with ϕ' still zero) (we denote the line $\zeta_1 = 0$ in the parameter space (ζ_1, ζ_2) by L_0; see Figure 12), or at $\zeta_2 = 0$ and $\zeta_1 < 0$ (see the half-line H_0 in Figure 12) through a pair of purely imaginary eigenvalues $\pm i\omega_0$, when a small amplitude periodic solution $r = r(t)$ appears with ϕ' still zero. This case will correspond to an SW solution. We mention that the $O(2)$-symmetry of the system forces both TW and SW solutions to appear simultaneously from the trivial solution.

The SS corresponds to solution of (5.6) *constant on the radial direction*, $r(t) = r_0$, and *with no orbital motion* $\phi' = 0$. This means that r_0 must satisfy the condition $\zeta_1 + Ar_0^2 = 0$ and that, obviously, it does not exist for all parameter values. In order to get an SS we need $A\zeta_1 < 0$ so that $r_0 = \sqrt{-\zeta_1/A}$. Moreover, $\phi' = 0$ implies $\phi(t) = \omega$, constant, and $z(t) = r_0\,e^{i\omega}$, $w(t) = z'(t) = 0$. The approximating formula (5.4) implies $U(x,t) = 2r_0\Phi_0\cos(k_0x + \omega) = 2\sqrt{-\frac{\zeta_1}{A}}\Phi_0\cos(k_0x + \omega)$, or up to a translation in space,

$$U(x,t) \approx 2\sqrt{-\frac{\zeta_1}{A}}\,\cos(k_0x)\,\Phi_0\,. \tag{5.7}$$

The SS pattern consists of oscillations with respect to the position in space x, and it is independent of time; therefore it forms stationary stripes (see Figure 19 for an example).

The linearization of (5.6) around $r = r_0$, $\phi' = 0$, e.g., take $r = r_0 + \xi$, $\phi' = 0 + \eta$ with ξ, η small, is the system of equations $\xi'' - (\zeta_2 + Mr_0^2)\,\xi' - 2Ar_0^2\,\xi = 0$, $\eta' - (\zeta_2 + Dr_0^2)\eta = 0$, i.e., $\xi'' - (\zeta_2 - \frac{M}{A}\zeta_1)\xi' + 2\zeta_1\xi = 0$ and $\eta' - (\zeta_2 - \frac{D}{A}\zeta_1)\eta = 0$. The only possible bifurcations that result in appearance/disappearance of time-independent solutions correspond to a zero eigenvalue. That can happen for $\zeta_1 = 0$ (we have already mentioned this case of a new SS branch solution) or for $A\zeta_2 = D\zeta_1$, $A\zeta_1 < 0$ (see the half-line L_m in Figure 12) when a new ϕ' constant and nonzero solution is created, say $\phi' = \omega_0$ (see the TW case).

The TW corresponds to a solution of (5.6) *constant on the radial direction*, $r(t) = r_0$, but *with orbital motion with constant angular frequency* $\phi' = \omega_0$. This means that r_0 and ω_0 must satisfy the conditions $\zeta_2 + Dr_0^2 = 0$ and $\omega_0^2 = -(\zeta_1 + Ar_0^2)$ and TW exists only in the parametric regime $D\zeta_2 < 0$, $\frac{A}{D}\zeta_2 - \zeta_1 > 0$. We have $r_0 = \sqrt{-\zeta_2/D}$ and $\omega_0 = \pm\sqrt{A\zeta_2/D - \zeta_1}$; then $z(t) = r_0\,e^{i(\omega_0t+\omega)}$, $w(t) = z'(t) = ir_0\omega_0\,e^{i(\omega_0t+\omega)}$ and from formula (5.4), the TW equation, up to a translation in space, is

$$U(x,t) \approx 2\sqrt{-\frac{\zeta_2}{D}}\;\mathrm{Re}\left[(\Phi_0 + i\omega_0\Phi_1)\,e^{i(\omega_0t+k_0x)}\right], \tag{5.8}$$

with

$$\omega_0 = \pm\sqrt{\frac{A}{D}\zeta_2 - \zeta_1}\,. \tag{5.9}$$

The TW solution changes with respect to time and position in space according to the TW coordinate $\xi = ct - x$, where $c = \omega_0/k_0$ is the wave velocity; therefore the pattern is formed by nonstationary stripes, i.e., stripes with finite slope (see Figure 16 for an example). Equation (5.8) shows that TW solutions break reflection symmetry but respect the symmetry to rotations.

The MW corresponds to a periodic solution $r(t)$ of (5.6) and a nonzero angular velocity ϕ'. This means that we have *oscillations in the radial direction* and *orbital motion* as well. The MWs bifurcate from a TW through another Hopf bifurcation that introduces a new frequency in the solution. Therefore the MW pattern is characterized by two different frequencies, one corresponding to the orbital motion and the other to radial oscillations (not shown in this paper).

The SW corresponds to a periodic solution $r(t)$ of (5.6) and $\phi' = 0$. This means that we have *oscillations in the radial direction* and *no orbital motion.* SWs can occur as oscillations about the trivial solution or about an SS. From (5.6), $r(t)$ satisfies the equation $r'' - r'\,(\zeta_2 + Mr^2) - r(\zeta_1 + Ar^2) = 0$ and $\phi = \omega$ is constant. Then $z(t) = r(t)\,e^{i\omega}$, $w(t) = z'(t) = r'(t)\,e^{i\omega}$, and (5.4) implies, up to a translation in space,

$$U(x,t) \approx 2\big[r(t)\,\Phi_0 + r'(t)\,\Phi_1\big]\;\cos(k_0 x)\,, \tag{5.10}$$

where $r(t)$ is the periodic solution of period, say, $2\pi/\omega_0$ of the ODE

$$r'' - r'\,(\zeta_2 + Mr^2) - r(\zeta_1 + Ar^2) = 0\,. \tag{5.11}$$

The SW solution oscillates with respect to time with frequency ω_0 for any fixed position in space and oscillates with respect to space with frequency k_0 for any fixed t (see Figure 17 for an example of the SW pattern). Equation (5.10) shows that SW solutions break rotation symmetry but respect the symmetry to reflection.

Remark 11. We summarized above the properties of possible patterns in a system with $O(2)$-symmetry. Let us describe now the type of bifurcation diagram [6] that we will need later in our study. It corresponds to $A < 0$ with $D < 0$, $M < 0$, and $0 < D/M < \frac{1}{2}$. The parameter plane (ζ_1, ζ_2) is divided into seven regions (see Figure 12) by the following curves: $L_0 : \zeta_1 = 0$, $H_0 : [\zeta_2 = 0,\ \zeta_1 < 0]$, $L_M : [A\zeta_2 = M\zeta_1,\ \zeta_1 > 0]$, $SL_S : [5A\zeta_2 = 4M\zeta_1,\ \zeta_1 > 0]$, $SN_{S2} : [A\zeta_2 \approx 0.74M\zeta_1,\ \zeta_1 > 0]$, $L_m : [A\zeta_2 = D\zeta_1,\ \zeta_1 > 0]$. A bifurcation producing SS solutions occurs along L_0, and a Hopf bifurcation, from the trivial solution T, of a TW and SW_1 occurs along H_0. By crossing L_M, SL_S, SN_{S2}, and L_m secondary bifurcations occur: along SN_{S2} we have a saddle-node for two SWs, SW_1 and SW_2; along L_M an SW oscillation SW_3 about a nontrivial SS bifurcates; then SW_3 and SW_2 undergo a global bifurcation and join smoothly to each other along SL_S; and TW bifurcates from an SS along L_m.

Since our goal is to study how the SS, TW, and SW solutions occur in the neural system (1.3) and how they interact, that is, how the patterns change as the parameters α and g vary about α^* and g^*, the next necessary step in the analysis is to construct the normal form for the double-zero bifurcation and determine its coefficients. That is the aim of the next section.

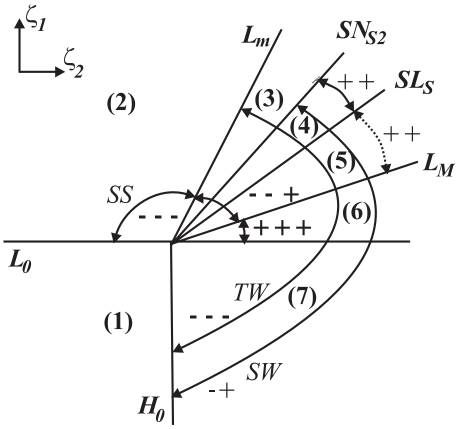


Figure 12. *The bifurcation diagram corresponding to system (5.5) with $A < 0$, $D < 0$, $M < 0$, and $0 < D/M < 1/2$.*

5.1. Double-zero bifurcation with $O(2)$-symmetry and pattern formation.

In this section we construct the normal form for the double-zero bifurcation with $O(2)$-symmetry for neural system (1.3) (see Appendix A for details) and delineate the regions in the parameter space that corresponds to different possible scenarios.

We obtain the following results.

Theorem 5.1. *For any positive τ, in the neighborhood of the bifurcation values $\alpha^* = \frac{1+1/\tau}{\hat{J}(k_0)}$ and $g^* = 1/\tau$, system (1.3) has the normal form (5.5) with $\zeta_1 = \frac{\alpha\,\hat{J}(k_0)-(g+1)}{\tau}$, $\zeta_2 = \alpha\,\hat{J}(k_0) - (1+\frac{1}{\tau})$, and the coefficients*

$$\begin{cases} A = \frac{1}{2\tau^2}[F'''(0) - 3F''(0)^2] + \frac{F''(0)^2}{\tau(\tau+1)} \cdot \left[\frac{\hat{J}(k_0)}{\hat{J}(k_0)-\hat{J}(0)} + \frac{\hat{J}(k_0)}{2\,[\,\hat{J}(k_0)-\hat{J}(2k_0)\,]}\right], \\ C = (\tau+1)A + \frac{F''(0)^2}{\tau(\tau+1)} \cdot \frac{\hat{J}(k_0)}{\hat{J}(k_0)-\hat{J}(0)}, \\ D = (\tau+1)A + \frac{F''(0)^2}{\tau(\tau+1)} \cdot \frac{\hat{J}(k_0)}{\hat{J}(k_0)-\hat{J}(2k_0)}. \end{cases}$$

Theorem 5.2. *In the hypotheses of Theorem 5.1, the SS, TW, and SW solutions that occur about the bifurcation point $\alpha^* = \frac{1+1/\tau}{\hat{J}(k_0)}$ and $g^* = 1/\tau$ are approximated by the following expressions:*

$$SS\ :\ u(x,t) = v(x,t) \approx \frac{2}{\tau}\cos(k_0 x)\sqrt{\frac{\alpha\,\hat{J}(k_0)-(g+1)}{(-A)}}\,,$$

$$TW\ :\ \begin{cases} u(x,t) \approx [\,\cos(\omega_0 t + k_0 x) - \tau\,\omega_0 \sin(\omega_0 t + k_0 x)\,]\sqrt{\frac{2[\alpha\,\hat{J}(k_0)-(1+1/\tau)]}{(-D)\,\tau}}\,, \\ v(x,t) \approx \cos(\omega_0 t + k_0 x)\sqrt{\frac{2[\alpha\,\hat{J}(k_0)-(1+1/\tau)]}{(-D)\,\tau}}\,, \end{cases}$$

$$SW\ :\ \begin{cases} u(x,t) \approx \frac{2}{\sqrt{\tau}}\,[\,r(t) + \tau\, r'(t)\,]\,\cos(k_0 x)\,, \\ v(x,t) \approx \frac{2}{\sqrt{\tau}}\, r(t)\,\cos(k_0 x)\,, \end{cases}$$

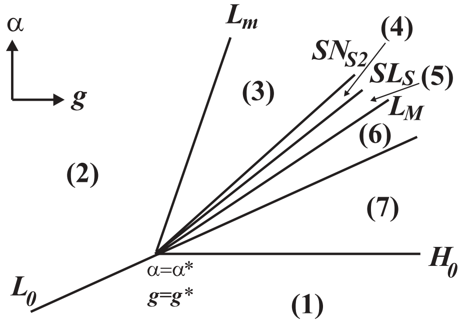


Figure 13. *The bifurcation diagram corresponding to system* (1.3) *about* (α^*, g^*) *when* $\theta = 0$ *in* F.

where $r(t)$ *is the periodic solution of*

$$r'' - r'\left[[\alpha\,\hat{J}(k_0) - (1 + 1/\tau)] + (2C + D)r^2\right] - r\left[\frac{\alpha\,\hat{J}(k_0) - (g+1)}{\tau} + Ar^2\right] = 0$$

and

$$\omega_0 = \pm\sqrt{\frac{A}{D}[\alpha\,\hat{J}(k_0) - (1 + 1/\tau)] - \frac{1}{\tau}[\alpha\,\hat{J}(k_0) - (g+1)]}\,.$$

Proof. The above formulas result directly from (5.7), (5.8), and (5.10) with Φ_0 and Φ_1 defined by (5.2) and ζ_1, ζ_2 as in Theorem 5.1. ■

Theorem 5.3. *Let us assume that the most unstable mode* k_0 *of system* (1.3) *satisfies conditions* (2.5), (2.6), *and at* k_0 *a double-zero eigenvalue occurs.*

If the firing-rate function F *is such that* $F(0) = 0$, $F'(0) > 0$, $F''(0) = 0$, *and* $F'''(0) < 0$, *then about* $\alpha^* = \frac{1+1/\tau}{\hat{J}(k_0)}$, $g^* = 1/\tau$, *system* (1.3) *has the bifurcation diagram from Figure* 13 *(equivalent to Figure* 12*). The curves that divide the parametric plane* (α, g) *into seven regions have the following equations:*

$$\begin{cases} L_0\ :\ \alpha\,\hat{J}(k_0) = g + 1\,, \\ H_0\ :\ \alpha\,\hat{J}(k_0) = 1 + 1/\tau\,,\ \ g > 1/\tau\,, \\ L_M\ :\ \alpha\,\hat{J}(k_0) = \frac{\tau+1}{2\tau+3}(3g+2)\,,\ \ g > 1/\tau\,, \\ SL_S\ :\ \alpha\,\hat{J}(k_0) = \frac{\tau+1}{7\tau+12}(12g+7)\,,\ \ g > 1/\tau\,, \\ SN_{S2}\ :\ \alpha\,\hat{J}(k_0) \approx \frac{\tau+1}{61\tau+111}(111g+61)\,,\ \ g > 1/\tau\,, \\ L_m\ :\ \alpha\,\hat{J}(k_0) = (\tau+1)g\,,\ \ g > 1/\tau\,. \end{cases} \tag{5.12}$$

Proof. Since $F''(0) = 0$ and $F'''(0) < 0$, we have $A < 0$, $C = D = (\tau+1)A < 0$ and then $M < 0$, $D/M = \frac{1}{3} \neq \frac{1}{2}, \frac{3}{5}, 0.7, 0.74, \frac{3}{4}, \frac{4}{5}, 1$ (the nondegeneracy conditions). This is exactly the case described by Figure 12. With the formulas provided by Theorem 5.1 we obtain immediately (5.12). ■

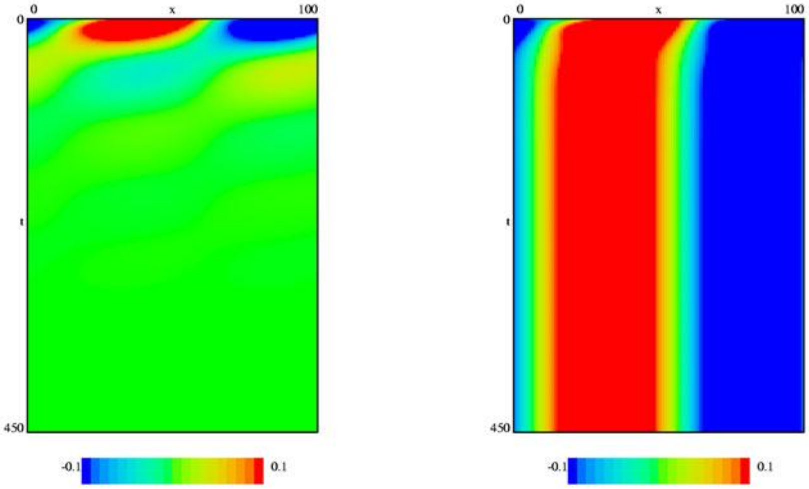


Figure 14. *Along the bifurcation line L_0 we have* (a) *the trivial solution T in region* (1)*, at $\alpha = 0.95$, $g = 0.2$, and* (b) *SS in region* (2)*, at $\alpha = 0.98$, $g = 0.2$. The same set of initial conditions Ic1 is used.*

5.2. Numerical results. We run the numerical simulations for the same hypotheses as in section 3.2. The full system (1.3) is simulated with synaptic coupling (1.7) that has the coefficients $a = -0.2$, $b = 2.5$, $c = 2$; the gain function F is chosen with $r = 3$ and $\theta = 0$, and the parameter τ is fixed at $\tau = 4$. The horizontal axis represents space, x, the vertical axis corresponds to time, t, and the variable $u(x, t)$ is plotted by the change in the level of color.

Therefore we have $F''(0) = 0$, $k_0 = 1$, $\hat{J}(0) = -0.2$, $\hat{J}(k_0) = 1.25$, $\hat{J}(2k_0) = 1$, and $\alpha^* = 1$, $g^* = 0.25$. The coefficients in the normal form are $A = -0.1406$, $C = D = -0.7031$, and then $M = -2.1094$, $D/M = \frac{1}{3}$. From Theorem 5.3 we obtain $L_0 : \alpha = \frac{4}{5}(g + 1)$, $H_0 : [\alpha = 1,\ g > 0.25]$, $L_M : [\alpha = \frac{12}{11}(g + \frac{2}{3}),\ g > 0.25]$, $SL_S : [\alpha = \frac{6}{5}(g + \frac{7}{12}),\ g > 0.25]$, $SN_{S2} : [\alpha = \frac{444}{355}(g + \frac{61}{111}),\ g > 0.25]$, $L_m : [\alpha = 4g,\ g > 0.25]$.

We choose parameters in different regions.

Along the bifurcation line L_0 we take $\alpha = 0.95$, $g = 0.2$ in region (1) and obtain the trivial solution T, and take $\alpha = 0.98$, $g = 0.2$ in region (2) and obtain the SS pattern (Figure 14). The same set of initial conditions, say Ic1, is considered in both cases. This set of initial conditions will be used later for other parameter values.

Along the bifurcation line H_0 let us take $\alpha = 0.98$, $g = 0.26$ in region (1) (Figure 15), and $\alpha = 1.004$, $g = 0.26$ in region (7). We consider two distinct sets of initial conditions, Ic1, the same as above, and Ic2. By crossing H_0 the patterns that bifurcate from T are different: TW for Ic1 (Figure 16), and SW for Ic2, but this is unstable (see Figures 17 and 18).

Along the bifurcation line L_m we consider $\alpha = 1.08$, $g = 0.26$ in region (2) with different initial conditions Ic1 and Ic2. An SS pattern is selected (Figure 19). At $\alpha = 1.03$, $g = 0.26$ in region (3) we obtain a TW pattern for Ic1, and an SS pattern for Ic2 (Figure 20). The SS pattern is unstable, as we see when we introduce in the system white noise scaled by a factor $nz = 0.001$ (Figure 21).

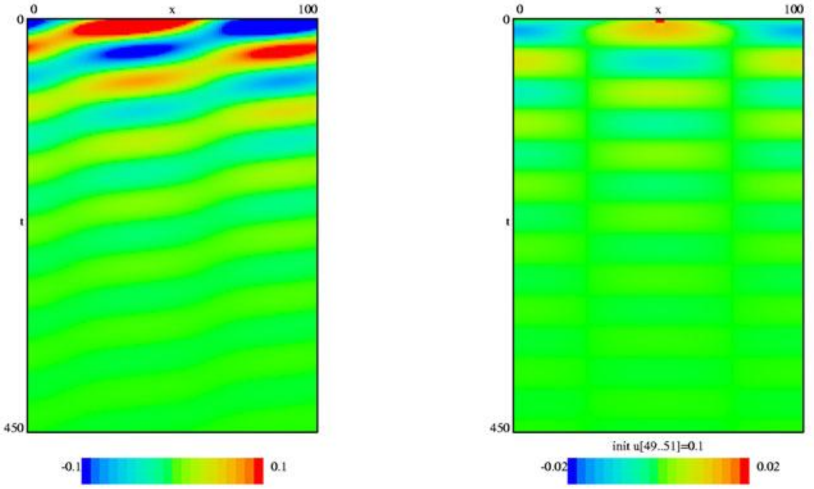


Figure 15. *At $\alpha = 0.98$, $g = 0.26$ in region* (1)*, close to the bifurcation line H_0, we obtain T for different sets of initial conditions.* (a) *Ic1 will give rise in region* (7) *to a TW.* (b) *Ic2 will give rise in region* (7) *to an unstable SW.*

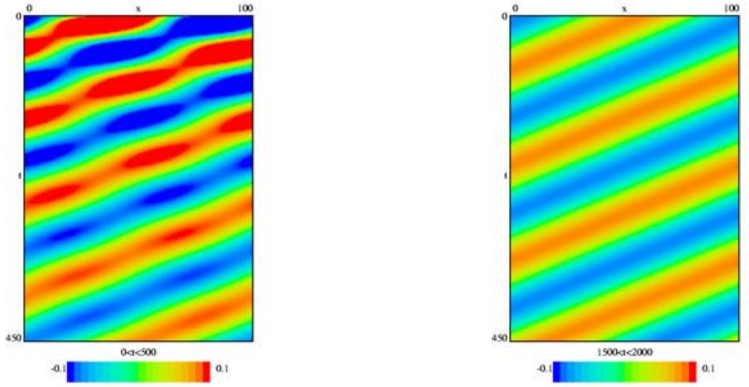


Figure 16. *The TW pattern obtained for Ic1, at $\alpha = 1.004$, $g = 0.26$ in region* (7).

We consider $\alpha = 1.0122$, $g = 0.26$ in region (4), between SN_{S2} and SL_S, and $\alpha = 1.01122$, $g = 0.26$ in region (5), between SL_S and L_M. For different initial conditions Ic1, Ic2, and Ic3, patterns such as TW, SW, and SS, respectively, can occur, but the last two are destabilized in time to a TW. We present, for example, the numerical results for $\alpha = 1.0122$, $g = 0.26$. Starting with initial condition Ic1 we obtain a TW pattern (Figure 22); starting with Ic2, an SW pattern is formed, but it destabilizes in time to a TW (Figure 23); starting with Ic3, an SS pattern is formed, but it destabilizes in time to a TW (Figure 24).

Similar pictures are obtained for $\alpha = 1.01122$, $g = 0.26$ in region (5).

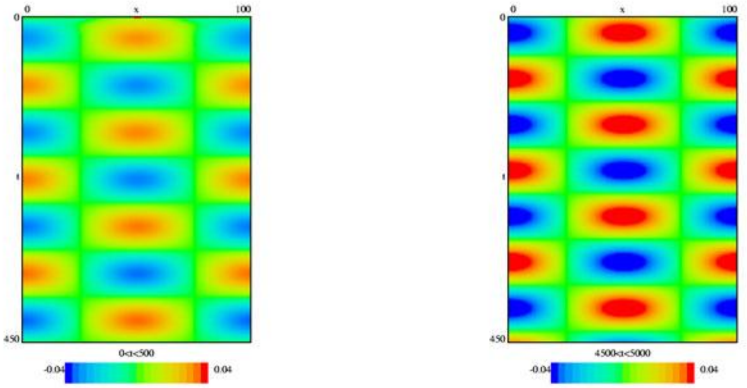


Figure 17. *The SW pattern obtained for Ic2, at $\alpha = 1.004$, $g = 0.26$ in region (7). This pattern is unstable.*

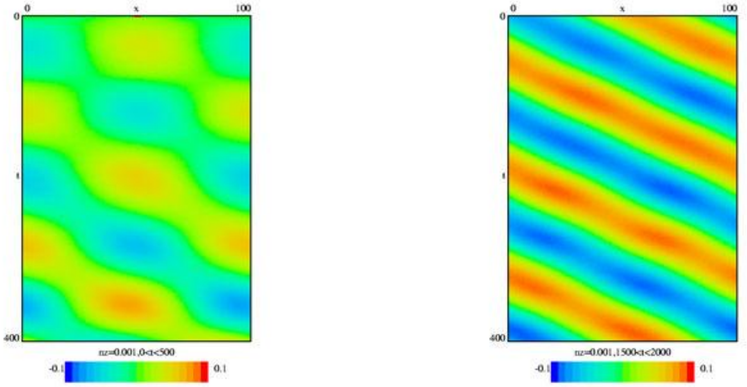


Figure 18. *The SW pattern obtained at $\alpha = 1.004$, $g = 0.26$ for Ic2 is destabilized to TW in the presence of noise.*

At $\alpha = 1.009$, $g = 0.26$ in region (6) the patterns that might occur are TW and SW, but always SW is destabilized in time to a TW (Figure 25).

Remark 12. There is a nice agreement between the theoretical and numerical results. Numerically it is impossible to detect a pattern that has all the corresponding eigenvalues positive, i.e., it is completely unstable (as one of the SWs in regions (4) and (5) in the bifurcation diagram, or the SS in region (6)). Nevertheless the other patterns that present stability at least in one direction can be visualized in the numerical simulation. Of course, eventually they will approach the only stable solution, i.e., TW.

Remark 13. We did not complete the analysis of system (1.3). A direction for future research is to investigate other possible cases (bifurcation diagrams) that might occur for different values of the parameter θ in the function F. We are especially interested in the case when the SW pattern is stable, and furthermore in the case that can give rise to an additional pattern not studied here, the MW pattern. These situations correspond to different kinds of

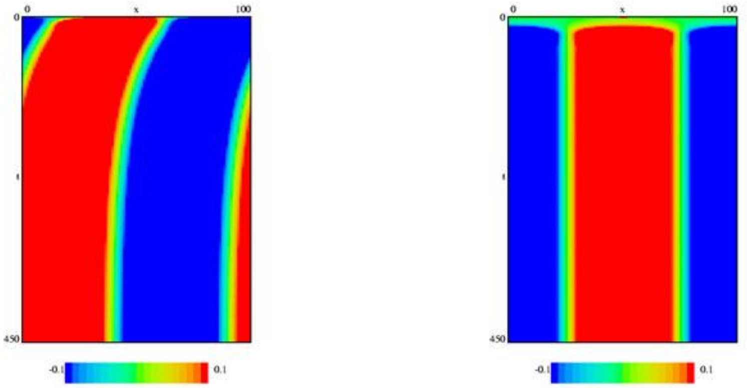


Figure 19. *The SS pattern obtained at* $\alpha = 1.08$*,* $g = 0.26$ *in region* (2)*, close to the bifurcation line* L_m *for initial conditions* (a) $Ic1$ *and* (b) $Ic2$*.*

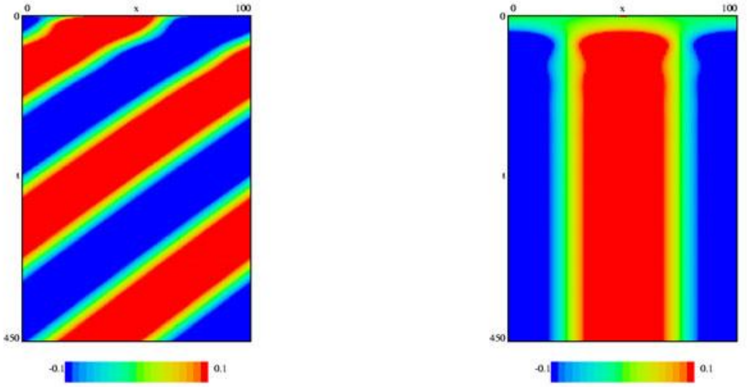


Figure 20. *At* $\alpha = 1.03$*,* $g = 0.26$ *in region* (3) *we obtain* (a) *TW for* $Ic1$ *and* (b) *SS for* $Ic2$ *(this pattern is unstable).*

bifurcation diagrams listed in [6]. Nevertheless, we have seen how to effect the transition from stationary patterns to TWs by varying the degree of adaptation.

6. Conclusions.

We have analyzed a rate model with nonlinear sigma-shaped gain function for two homogeneous populations of neurons, one excitatory that displays adaptation, and one inhibitory. The coupling is characterized by local excitation and long range (lateral) inhibition, and the adaptation is assumed to be linear. When the strength of adaptation is sufficiently large (or the adaptation is slow enough) temporal oscillations occur in the system. In general they form TWs, but we were able to show that it is possible, when the threshold is sufficiently high, to obtain also SWs. These are spatial oscillations with frequency k_0 and temporal oscillations with frequency ω_0 that can be computed as functions of parameters τ, g, and the strength of the coupling α. Numerical simulations indicate that for a fixed adaptation time constant τ, the SW pattern occurs for an intermediate value of the strength g of

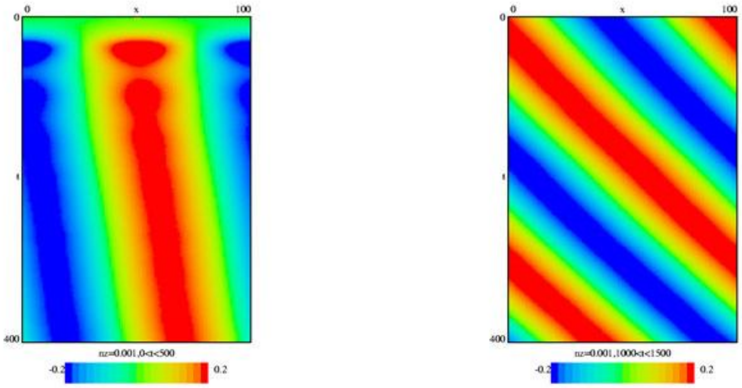


Figure 21. *The SS pattern obtained at $\alpha = 1.03$, $g = 0.26$ for Ic2 is destabilized to TW in the presence of noise.*

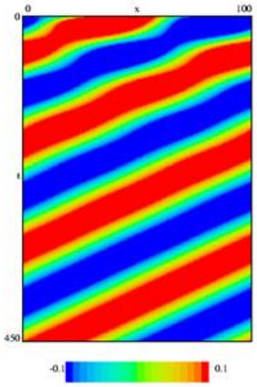


Figure 22. *At $\alpha = 1.0122$, $g = 0.26$ in region (4), starting with Ic1 initial conditions a TW pattern is formed.*

adaptation, in a relatively small regime. When g is increased, the local activity is disrupted and starts to travel along the network, resulting in a TW pattern. Our condition for delineating the onset of stationary versus time-dependent patterns is identical to that in Hansel and Sompolinsky; they distinguish the strong and weak adaptation cases in their equations (13.80), (13.81).

We have also investigated the transition between stationary patterns and spatio-temporal patterns in the neural network, therefore explaining the patterns found in the numerical simulations of the full model. We did not complete the analysis of system (1.3). The general theory predicts, under certain conditions, the existence of a different spatio-temporal pattern, MWs characterized by two different temporal frequencies in addition to the spatial frequency k_0. The question of whether or not there are parameters regimes in the neural models in which there are MWs remains open.

We have made an assumption that there is slow negative feedback in the form of additive

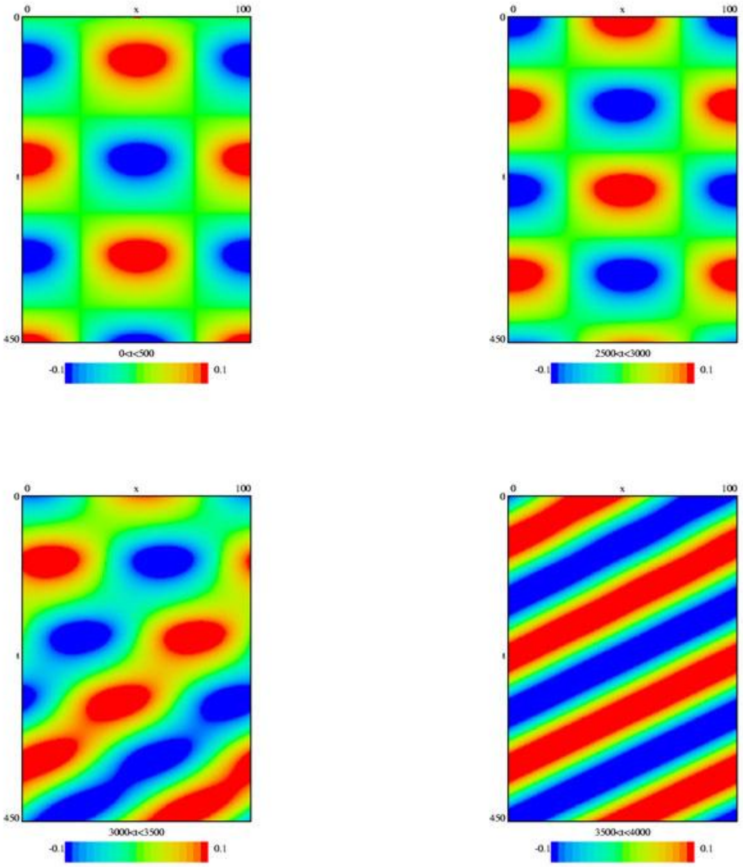


Figure 23. *At $\alpha = 1.0122$, $g = 0.26$ in region (4), starting with Ic2 initial conditions, an SW pattern is formed. Nevertheless it is destabilized in time to a TW.*

adaptation. However, this is not the only way to get slow modulation of excitation. Indeed, an alternate model could use synaptic depression. If we assume that such depression occurs for excitatory-excitatory and excitatory-inhibitory synapses, then in (1.3) we delete the $-gv$ term and replace α by αv. The v equation has the form

$$\tau \frac{dv}{dt} = q(u) - v,$$

where $q(u)$ is a monotone decreasing function of u. Thus, if u goes up, then v tends to zero corresponding to the depression of the synaptic activation. For low values of u, $q(u)$ tends to 1 so that the network is fully poised to fire. The linear stability analysis will be more complicated as will the computation of the normal forms, but we expect little change in the qualitative behavior. Other slow negative feedback mechanisms such as accumulation of intracellular sodium (leading to a reduction in the excitability) could also be modeled and would result in qualitatively similar behavior and the same basic bifurcation picture.

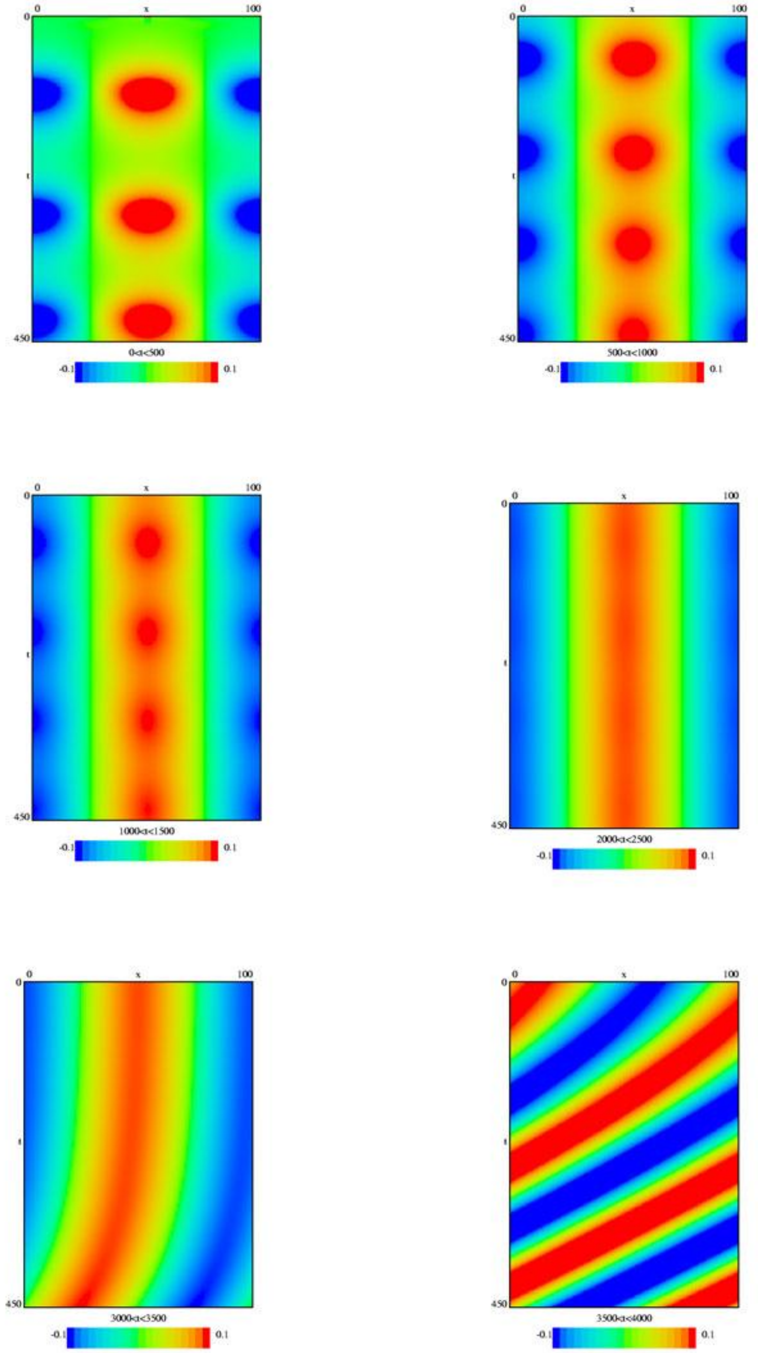


Figure 24. *At $\alpha = 1.0122$, $g = 0.26$ in region (4), starting with Ic3 initial conditions, an SS pattern is formed. Nevertheless it is destabilized in time to a TW.*

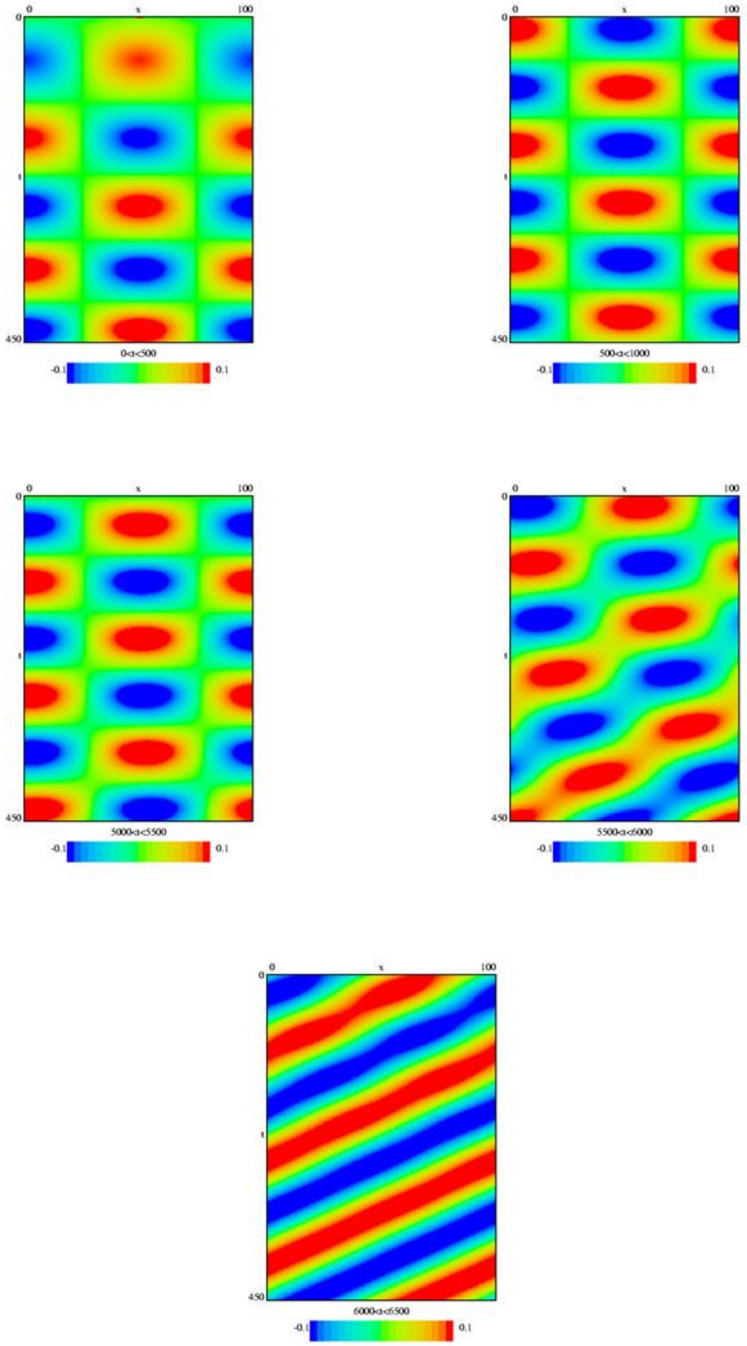


Figure 25. *At $\alpha = 1.009$, $g = 0.26$ in region (6), an SW pattern may form, but it destabilizes in time to a TW.*

We close with some speculations on the possible functional implications of the coalescence of the Hopf bifurcation and the SS bifurcation (that is, the Takens–Bogdanov bifurcation). We first point out that the original motivation of the HS model was to explain nonlinear amplification of orientation tuning. Assume that the domain is the circle so that a point x represents a specific orientation. Suppose that the time constant τ of the adaptation is small so that adaptation works quickly. Then, as we saw above, the only bifurcation is to stationary spatial patterns. As in section 3.3, let $k = 1$ be the critical wavenumber. This means that there will be a local peak at a point x_m; weak inputs will move that peak [8] so that it is pinned at a specific orientation. Thus, the network amplifies weak inputs to produce a large nonlinear response to a specific orientation. This is the *normal* state of the network. In various pathological situations, however, the network can be damaged so that this behavior is disrupted. In a recent paper, Prole, Lima, and Marrion showed that acidosis (increased pH) that occurs during epilepsy or stroke increased the time constant of both activation and inactivation of the slow voltage-dependent potassium current K_m [21]. This current is one of several that are responsible for frequency *adaptation* in cortical neurons. In the context of our simple model, this is equivalent to increasing τ_A, the time constant of adaptation. As we have shown, increasing τ_A leads to a transition from the zero eigenvalue stationary-state case to the oscillatory Hopf case. Waves have been associated with epileptic behavior in previous models (see, e.g., [20]).

The present model is for either a ring or a line of one-dimensional cortex. The actual cortex is better represented as a sheet. In this case, the dimensions of the null-space of the resulting linearized system can be much larger than studied here. For example, in the stationary pattern case, there can be stripe patterns and square patterns which bifurcate from rest. The selection between these patterns depends on the nonlinear terms in much the same way as the selection between SWs and TWs in the present paper. Two-dimensional models of cortical networks were used to explain the patterns observed during visual hallucinations [12]. Recently Bressloff and his collaborators [1] have extended the two-dimensional models to incorporate spatial connectivity and orientation selectivity in these models. Bifurcation methods should remain an important technique for the analysis of patterns in increasingly more realistic neural models.

Appendix A.

Hopf bifurcation and pattern formation. In the case of a pair of purely imaginary eigenvalues, system (2.1) can be written in the equivalent form

$$L_0 U = (\alpha - \alpha^*)\,(J * u\,,\,0)^T + B(U,U) + C(U,U,U) + \cdots \tag{A.1}$$

with

$$B(U,U) = \left(\frac{F''(0)}{2} (\alpha J * u - gv)^2, 0 \right)^T, \; C(U,U,U) = \left(\frac{F'''(0)}{6} (\alpha J * u - gv)^3, 0 \right)^T.$$

A good scaling for the bifurcation parameter α and the solution U we are seeking is $\alpha - \alpha^* = \epsilon^2 \gamma$, $\gamma \in \mathbb{R}$, and

$$\begin{aligned} U(x,t) &= \epsilon\, U_0(x,t) + \epsilon^2\, U_1(x,t) + \epsilon^3\, U_2(x,t) + \cdots \\ &= \epsilon \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \epsilon^2 \begin{pmatrix} u_1 \\ v_1 \end{pmatrix} + \epsilon^3 \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} + \cdots . \end{aligned} \tag{A.2}$$

With the notation $\boldsymbol{E} = (1,0)^T$, (A.1) and (A.2) imply

(A.3)

$$\epsilon L_0 U_0 + \epsilon^2 L_0 U_1 + \epsilon^3 L_0 U_2 + \mathcal{O}(\epsilon^4) = \epsilon^2 \boldsymbol{E}\, \frac{F''(0)}{2} [\alpha^*\, J * u_0 - g v_0]^2$$
$$+ \epsilon^3 \boldsymbol{E} \left[\gamma\, (J * u_0) + F''(0) [\alpha^*\, J * u_0 - g v_0][\alpha^*\, J * u_1 - g v_1] + \frac{F'''(0)}{6} [\alpha^*\, J * u_0 - g v_0]^3 \right] + \mathcal{O}(\epsilon^4);$$

therefore the first equation to be solved is $L_0 U_0 = \mathbf{0}$.

The nullspace of L_0 corresponding to the center manifold is four-dimensional with the basis $\left\{\Phi_0\, e^{i(\omega_0 t \pm k_0 x)}, \overline{\Phi}_0\, e^{-i(\omega_0 t \pm k_0 x)}\right\}$, and U_0 can be written as

$$U_0 = z_1\, \Phi_0\, e^{i(\omega_0 t + k_0 x)} + w_1\, \Phi_0\, e^{i(\omega_0 t - k_0 x)} + \overline{z}_1\, \overline{\Phi}_0\, e^{-i(\omega_0 t + k_0 x)} + \overline{w}_1\, \overline{\Phi}_0\, e^{-i(\omega_0 t - k_0 x)} .$$

Since in (A.3), $L_0 U_0 = \mathcal{O}(\epsilon)$, z_1 and w_1 are ϵ-dependent and we can write them as $z_1 = z_1(T)$ and $w_1 = w_1(T)$ with $T = \epsilon^2 t$ a slow time. These imply the singular perturbation expansions

$$\begin{aligned} z_1(T) &= z_1(T)_{|\epsilon=0} + z_1'(T)_{|\epsilon=0}\, \epsilon^2 t + \frac{1}{2} z_1''(T)_{|\epsilon=0}\, (\epsilon^2 t)^2 + \cdots , \\ w_1(T) &= w_1(T)_{|\epsilon=0} + w_1'(T)_{|\epsilon=0}\, \epsilon^2 t + \frac{1}{2} w_1''(T)_{|\epsilon=0}\, (\epsilon^2 t)^2 + \cdots . \end{aligned}$$

For simplicity we introduce the notation $z_2 = z_1(T)_{|\epsilon=0}$, $\dot{z}_2 = z_1'(T)_{|\epsilon=0}$, $w_2 = w_1(T)_{|\epsilon=0}$, $\dot{w}_2 = w_1'(T)_{|\epsilon=0}$. The dot shows the derivation of z_2 and w_2 with respect to the slow time T.

We obtain then $z_1 = z_2 + \dot{z}_2\, \epsilon^2 t + \mathcal{O}(\epsilon^4)$, $w_1 = w_2 + \dot{w}_2\, \epsilon^2 t + \mathcal{O}(\epsilon^4)$, and

$$\begin{aligned} U_0 = \Big[& z_2\, \Phi_0\, e^{i(\omega_0 t + k_0 x)} + w_2\, \Phi_0\, e^{i(\omega_0 t - k_0 x)} \\ & + \overline{z}_2\, \overline{\Phi}_0\, e^{-i(\omega_0 t + k_0 x)} + \overline{w}_2\, \overline{\Phi}_0\, e^{-i(\omega_0 t - k_0 x)} \Big] + \mathcal{O}(\epsilon^2) . \end{aligned}$$

Remark 14. The calculation of the normal form is cumbersome (see [5] for details of proofs). The normal form results by solving for U_1 and U_2 in the corresponding functional equations implied by (A.3), and it is

$$\begin{cases} \dot{z}_2 = z_2\, (\, \tilde{a} + b\, z_2\, \overline{z}_2 + c\, w_2\, \overline{w}_2\,) , \\ \dot{w}_2 = w_2\, (\, \tilde{a} + b\, w_2\, \overline{w}_2 + c\, z_2\, \overline{z}_2\,) , \end{cases} \tag{A.4}$$

with $\tilde{a} = a/\epsilon^2$, $a = \hat{J}(k_0)(\alpha - \alpha^*)\big(\frac{1}{2} - i\frac{1}{2\sqrt{g\tau - 1}}\big)$, and b, c constants [5]. We notice that, as a result of the scaling $z(t) = \epsilon\, z_2(T)$, $w(t) = \epsilon\, w_2(T)$ with $T = \epsilon^2 t$ (therefore $z' = dz/dt = \epsilon^3 \dot{z}_2$

and $w' = dw/dt = \epsilon^3 \dot{w}_2$), system (A.4) takes exactly the form (3.5) and, indeed, the linear approximation of $U(x,t)$ is (3.4).

Normal form for the pitchfork bifurcation. We fix the values of τ and g such that $g < 1/\tau$ and take α as the bifurcation parameter. The bifurcation value around which we will consider the singular perturbation analysis is α^* from (4.1), and, in the following, the operator L_0 and the matrix $\hat{L}(k)$ are evaluated at $\alpha = \alpha^*$.

The singular perturbation expansion for the parameter α and solution $U(x,t)$ follows exactly the same steps as in (A.1), (A.2), and (A.3). The nullspace of L_0 corresponding to the center manifold is now only two-dimensional and has the basis $\left\{\Phi_0\, e^{\pm ik_0x}\right\}$, so U_0 can be written as

$$U_0 = z_1\, \Phi_0\, e^{ik_0x} + \overline{z}_1\, \Phi_0\, e^{-ik_0x}$$

with z_1 depending on the slow time $T = \epsilon^2 t$, that is, $z_1 = z_1(T)$.

As in the previous subsection, we can write the singular perturbation expansion of z_1 as $z_1 = z_2 + \dot{z}_2\, \epsilon^2 t + \mathcal{O}(\epsilon^4)$. Then the equation that defines U_1 becomes

$$L_0 U_1 = \frac{F''(0)}{2} \boldsymbol{E} \left[z_2^2\, e^{2ik_0x} + \overline{z}_2^2\, e^{-2ik_0x} + 2 z_2\, \overline{z}_2 \right].$$

Therefore U_1 takes the form

$$\begin{aligned} U_1 &= \xi_1\, z_1^2\, e^{2ik_0x} + \xi_1\, \overline{z}_1^2\, e^{-2ik_0x} + 2\xi_2\, z_1\, \overline{z}_1 \\ &= \xi_1\, z_2^2\, e^{2ik_0x} + \xi_1\, \overline{z}_2^2\, e^{-2ik_0x} + 2\xi_2\, z_2\, \overline{z}_2 + \mathcal{O}(\epsilon^2) \end{aligned}$$

with ξ_1, ξ_2 defined by $\left[-\hat{L}(2k_0)\xi_1\right] = \frac{F''(0)}{2}\boldsymbol{E}$, $\left[-\hat{L}(0)\xi_2\right] = \frac{F''(0)}{2}\boldsymbol{E}$. The normal form corresponding to a zero eigenvalue results by solving the functional equation for U_2, and it is

$$\dot{z}_2 = z_2(\, \tilde{\eta}_1 + \Lambda z_2\, \overline{z}_2\,) \tag{A.5}$$

with $\tilde{\eta}_1 = \eta_1/\epsilon^2$, $\eta_1 = \frac{\hat{J}(k_0)(\alpha-\alpha^*)}{1-g\tau}$, and Λ defined by (4.3).

We notice that $U = \epsilon U_0 + \mathcal{O}(\epsilon^2) = \epsilon\, z_2\, \Phi_0\, e^{ik_0x} + \epsilon\, \overline{z}_2\, \Phi_0\, e^{-ik_0x} + \mathcal{O}(\epsilon^2)$. With the choice of $z(t) = \epsilon\, z_2(T)$, $T = \epsilon^2 t$, (A.5) is equivalent to

$$z' = \frac{\hat{J}(k_0)}{1 - g\tau}(\alpha - \alpha^*) z + \Lambda |z|^2 z\,,$$

which is the normal form at zero eigenvalue written in the original parameters.

Normal form for the double-zero bifurcation with $O(2)$-symmetry. In the case of a double-zero eigenvalue, we choose the singular perturbation expansion for α, g and the solution U as $\alpha - \alpha^* = \epsilon^2\, \gamma$, $g - g^* = \epsilon^2\, \eta$, $\gamma, \eta \in \mathbb{R}$,

$$U(x,t) = \epsilon\, U_0(x,t) + \epsilon^2\, U_1(x,t) + \epsilon^3\, U_2(x,t) + \epsilon^4\, U_3(x,t) + \cdots. \tag{A.6}$$

The system equivalent to (2.1) is then

$$L_0U = (\alpha - \alpha^*)\, L_1U + (g - g^*)\, L_2U + B(U,U) + C(U,U,U) + Q(U,U,U,U) + \cdots\,, \tag{A.7}$$

where $B(U,U)$, $C(U,U,U)$, $Q(U,U,U,U)$ represent the quadratic, cubic, and fourth order terms, L_0 is defined according to Remark 10, and

$$L_1 = \begin{pmatrix} J*(\cdot) & 0 \\ 0 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

With the notation $\boldsymbol{E}$ for the unit vector $(1,0)^T$, (A.7) becomes

$$\begin{aligned}
&\epsilon L_0U_0 + \epsilon^2 \left[L_0U_1 - \frac{F''(0)}{2}[\alpha^*\, J*u_0 - g^*v_0]^2 \boldsymbol{E} \right] + \epsilon^3 [\, L_0U_2 - \gamma\, L_1U_0 - \eta\, L_2U_0\,] \\
&\qquad + \epsilon^4 [\, L_0U_3 - \gamma\, L_1U_1 - \eta\, L_2U_1\,] \\
&= \epsilon^3 \left[F''(0)[\alpha^*\, J*u_0 - g^*v_0][\alpha^*\, J*u_1 - g^*v_1] + \frac{F'''(0)}{6}[\alpha^*\, J*u_0 - gv_0]^3 \right] \boldsymbol{E} \\
&\qquad + \epsilon^4 \Bigg[\frac{F''(0)}{2}[\alpha^*\, J*u_1 - g^*v_1]^2 + F''(0)[\alpha^*\, J*u_0 - g^*v_0][\alpha^*\, J*u_2 - g^*v_2 \\
&\qquad\qquad + \gamma\, J*u_0 - \eta\, v_0] + \frac{F'''(0)}{2}[\alpha^*\, J*u_0 - gv_0]^2[\alpha^*\, J*u_1 - gv_1] \\
&\qquad\qquad + \frac{F^{(4)}(0)}{24}[\alpha^*\, J*u_0 - gv_0]^4 \Bigg] \boldsymbol{E} + \mathcal{O}(\epsilon^5)\,.
\end{aligned} \tag{A.8}$$

In order to construct the normal form we need to identify the functional equations that U_0, U_1, U_2, and U_3 satisfy and then solve for them [5].

The nullspace of L_0 corresponding to the center manifold is now only two-dimensional with the basis $\left\{\Phi_0\, e^{\pm ik_0x}\right\}$, where Φ_0 is the real vector defined in (5.2). As a consequence, U_0 can be written as

$$U_0 = \left(z_1\, e^{ik_0x} + \overline{z}_1\, e^{-ik_0x}\right) \Phi_0$$

with $z_1 = z_1(T)$, where $T = \epsilon t$ is the appropriate slow time. Then

$$z_1(T) = z_1(T)_{|\epsilon=0} + z_1'(T)_{|\epsilon=0}\, \epsilon t + \frac{1}{2} z_1''(T)_{|\epsilon=0}\, (\epsilon t)^2 + \frac{1}{6} z_1'''(T)_{|\epsilon=0}\, (\epsilon t)^3 + \mathcal{O}(\epsilon^4)$$

or, with notation $z_2 = z_1(T)_{|\epsilon=0}$, $\dot{z}_2 = z_1'(T)_{|\epsilon=0}$,

$$z_1 = z_2 + \dot{z}_2\, \epsilon t + \frac{1}{2}\, \ddot{z}_2\, (\epsilon t)^2 + \frac{1}{6}\, \dddot{z}_2\, (\epsilon t)^3 + \mathcal{O}(\epsilon^4)\,.$$

Therefore $U_0 = \left(z_2\, e^{ik_0x} + \overline{z}_2\, e^{-ik_0x}\right) \Phi_0 + \mathcal{O}(\epsilon)$.

The equation that defines U_1 reads as

$$L_0U_1 = -\left[\dot{z}_2\, e^{ik_0x} + \dot{\overline{z}}_2\, e^{-ik_0x}\right] \Phi_0 + \frac{F''(0)}{2\tau} \boldsymbol{E} \left[z_2^2\, e^{2ik_0x} + \overline{z}_2^2\, e^{-2ik_0x} + 2 z_2\, \overline{z}_2\right]$$

and U_1 can be constructed as

$$U_1 = \left[w_1\, e^{ik_0x} + \overline{w}_1\, e^{-ik_0x}\right] \Phi_1 + z_1^2\xi_1\, e^{2ik_0x} + \overline{z}_1^2\xi_1\, e^{-2ik_0x} + 2z_1\,\overline{z}_1\xi_2$$

with ξ_1, ξ_2 real vectors and $w_1 = w_1(T)$, or, similar to the singular perturbation expansion of z_1,

$$w_1 = w_2 + \dot{w}_2\,\epsilon t + \frac{1}{2}\ddot{w}_2\,\epsilon^2 t^2 + \mathcal{O}(\epsilon^3)\,.$$

The first equation of the normal form is obtained by solving for U_1 and it is $\dot{z}_2 = w_2$.

The next two steps consist of finding U_2 as

$$U_2 = \left[z_1 w_1 e^{2ik_0x} + \overline{z}_1\,\overline{w}_1 e^{-2ik_0x}\right]\beta_1 + [z_1\,\overline{w}_1 + \overline{z}_1\,w_1]\beta_2 + \left[z_1^3 e^{3ik_0x} + \overline{z}_1^3 e^{-3ik_0x}\right]\beta_3$$

with $\beta_1, \beta_2, \beta_3$ real vectors to be computed. Then we need to solve for U_3 [5].

As a consequence we obtain the normal form

$$\text{(A.9)}\qquad \begin{cases} \dot{z}_2 = w_2\,, \\ \dot{w}_2 = \frac{\gamma\hat{J}(k_0)-\eta}{\tau} z_2 + A|z_2|^2 z_2 \\ \qquad\qquad + \epsilon\{\gamma\hat{J}(k_0)w_2 + Cz_2[\overline{z}_2 w_2 + z_2\overline{w}_2] + D|z_2|^2 w_2\} + \mathcal{O}(\epsilon^2)\,, \end{cases}$$

where $\gamma = (\alpha - \alpha^*)/\epsilon^2$ and $\eta = (g - g^*)/\epsilon^2$.

With the proper scaling $z(t) = \epsilon z_2(T)$, $w(t) = \epsilon^2 w_2(T)$, we have $z' = \epsilon^2 \dot{z}_2$, $w' = \epsilon^3 \dot{w}_2$, and (A.9) is equivalent to the normal form (5.5).

Remark 15. If we consider the solution U of the nonlinear system approximated only by its projection on the generalized eigenspace, we have

$$\begin{aligned} U(x,t) &\approx 2\epsilon\,\Phi_0\,\mathrm{Re}\left[z_2(T)\,e^{ik_0x}\right] + 2\epsilon^2\,\Phi_1\,\mathrm{Re}\left[w_2(T)\,e^{ik_0x}\right] \\ &\approx 2\,\Phi_0\,\mathrm{Re}\left[z(t)\,e^{ik_0x}\right] + 2\,\Phi_1\,\mathrm{Re}\left[w(t)\,e^{ik_0x}\right] \end{aligned}$$

and this is exactly the formula (5.4).

Appendix B. Additional material. The transitions from an unstable to a new stable pattern as a parameter is changed are shown in the following files. In each case, a pattern that was stable for one set of parameters is used as an initial condition for the new parameter. Often the patterns go through several intermediate transient states before settling into a final stable state. The main parameter that is varied is the threshold since this has no effect on the linear stability but rather affects the selection between patterns.

Animation 1. The initial data and parameter values are as in Figure 4. We start with $\theta = 0.3$ and initial conditions close to a TW which is stable for $\theta = 0.0$. The pattern evolves and stabilizes to an SW.

Animation 2. The last conditions in Animation 1 are chosen as initial conditions in Animation 2 and the value of parameter θ is changed to $\theta = 0$. The resulting stable pattern is now a TW.

Animation 4. This pattern is taken from the results of section 5.2 (Figure 24). An SW pattern is selected and then destabilized to a TW. Note the initial transition to a stationary pattern before switching to a TW.

Animation 5. The results presented in Figure 25 are shown here. An SW pattern is selected and destabilized to a TW.

Acknowledgment. Rodica Curtu thanks the University of Pittsburgh for support in the form of an Andrew Mellon Predoctoral Fellowship for her thesis work.

REFERENCES

[1] 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*, Philos. Trans. Roy. Soc. Lond. Ser. B, 40 (2001), pp. 299–330.

[2] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Periodicity and directionality in the propagation of epileptiform discharges across neocortex*, J. Neurophysiol., 60 (1988), pp. 1695–1713.

[3] B. W. Connors and M. J. Gutnick, *Intrinsic firing patterns of diverse neocortical neurons*, Trends Neurosci., 13 (1990), pp. 99–104.

[4] B. W. Connors, M. J. Gutnick, and D. A. Prince, *Electrophysiological properties of neocortical neurons in vitro*, J. Neurophysiol., 48 (1982), pp. 1302–1320.

[5] R. Curtu, *Waves and Oscillations in Model Neuronal Networks*, Ph.D. Dissertation, Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, 2003.

[6] G. Dangelmayr and E. Knobloch, *The Takens-Bogdanov bifurcation with $O(2)$-symmetry*, Philos. Trans. Roy. Soc. Lond. Ser. A, 322 (1987), pp. 243–279.

[7] B. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, SIAM, Philadelphia, 2002.

[8] G. B. Ermentrout, *Neural networks as spatio-temporal pattern-forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[9] B. Ermentrout, *Stripes or spots? Nonlinear effects in bifurcation of reaction-diffusion equations on the square*, Proc. Roy. Soc. Lond. Ser. A, 434 (1991), pp. 413–417.

[10] G. B. Ermentrout, *Symmetry Breaking in Homogeneous, Isotropic, Stationary Neuronal Nets*, Ph.D. Dissertation, The University of Chicago, Chicago, IL, 1979.

[11] G. B. Ermentrout, J. Campbell, and G. Oster, *A model for shell patterns based on neural activity*, The Veliger, 28 (1986), pp. 369–388.

[12] G. B. Ermentrout and J. Cowan, *A mathematical theory of visual hallucination patterns*, Biol. Cybern., 34 (1979), pp. 137–150.

[13] G. B. Ermentrout and D. Kleinfeld, *Traveling electrical waves in cortex: Insights from phase dynamics and speculation on a computational role*, Neuron, 29 (2001), pp. 33–44.

[14] M. J. Gutnick, B. W. Connors, and D. A. Prince, *Mechanisms of neocortical epileptogenesis in vitro*, J. Neurophysiol., 48 (1982), pp. 1321–1335.

[15] D. Hansel and H. Sompolinsky, *Modeling feature selectivity in local cortical circuits*, in Methods in Neuronal Modeling: From Ions to Networks, 2nd ed., C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1998.

[16] D. Hubel and T. Wiesel, *Receptive fields and functional architecture of monkey striate cortex*, J. Physiol., 195 (1968), pp. 215–243.

[17] E. Knobloch and J. de Luca, *Amplitude equations for travelling wave convection*, Nonlinearity, 3 (1990), pp. 975–980.

[18] C. R. Laing and C. C. Chow, *A spiking neuron model for binocular rivalry*, J. Comp. Neurosci., 12 (2002), pp. 39–53.

[19] J. D. Murray, *Mathematical Biology*, 2nd ed., Springer, New York, 1998.

[20] 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.

[21] D. L. Prole, P. A. Lima, and N. V. Marrion, *Mechanisms underlying modulation of neuronal KCNQ2/KCNQ3 potassium channels by extracellular protons*, J. Gen. Physiol., 122 (2003), pp. 775–793.

[22] N. V. Swindale, *A model for the formation of ocular dominance stripes*, Proc. Roy. Soc. Lond. Ser. B, 208 (1980), pp. 243–264.

[23] H. R. Wilson, *Spikes, Decisions, and Actions*, Oxford University Press, Oxford, UK, 1999.

[24] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.

[25] X. Xie, R. H. R. Hahnloser, and H. S. Seung, *Double-ring network model of the head-direction system*, Phys. Rev. E, 66 (2002), 041902.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 3, pp. 232–260

A Fast Method for Approximating Invariant Manifolds[*]

John Guckenheimer[†] and Alexander Vladimirsky[†]

Abstract. The task of constructing higher-dimensional invariant manifolds for dynamical systems can be computationally expensive. We demonstrate that this problem can be locally reduced to solving a system of quasi-linear PDEs, which can be efficiently solved in an Eulerian framework. We construct a fast numerical method for solving the resulting system of discretized nonlinear equations. The efficiency stems from decoupling the system and ordering the computations to take advantage of the direction of information flow. We illustrate our approach by constructing two-dimensional invariant manifolds of hyperbolic equilibria in $\boldsymbol{R}^3$ and $\boldsymbol{R}^4$.

Key words. global stable and unstable manifolds, numerical method, ordered upwind

AMS subject classifications. 37D10, 37M20, 35F30, 65N22

DOI. 10.1137/030600179

1. Introduction.

Invariant manifolds are important in many application areas. In the context of dynamical systems theory, stable and unstable manifolds are fundamental geometric structures. They partition phase-spaces into sets of points with the same forward and backward limit sets. We cite the following three ways in which problems involving stable and unstable manifolds of equilibria arise:

1. Stable and unstable manifolds play a role in global bifurcation. Homoclinic and heteroclinic bifurcations occur at nontransverse intersections of stable and unstable manifolds. For example, a particular global bifurcation of the Kuramoto–Sivashinsky equation is examined in [19] using the method for approximating invariant manifolds developed in [18].
2. In studying the structure of weak shock waves for hyperbolic systems of conservation laws, the admissibility of a traveling-wave ansatz depends on the existence of a heteroclinic orbit connecting the left-state/right-state equilibria; see, for example, [35]. Such an orbit exists if the stable manifold of u_r intersects the unstable manifold of u_l.
3. For systems with multiple attractors whose basins cover all but the set of measure zero, a basin boundary can often be obtained from the stable manifolds of equilibria with a single unstable direction. Such delineation of basins is an important practical task. For example, transient stability analysis of power systems deals with the stability properties after an *event disturbance* modeled as a time-localized (*fault-on*) change in the vector field. The key test is to determine if a fault-on trajectory ends up inside the desired stability region of the postfault system [2]. On the other hand, in designing

[*] Received by the editors April 28, 2003; accepted for publication (in revised form) by M. Golubitsky February 10, 2004; published electronically July 6, 2004.
http://www.siam.org/journals/siads/3-3/60017.html

[†] Department of Mathematics, Cornell University, Ithaca, NY 14853 (gucken@cam.cornell.edu , vlad@math.cornell.edu). The first author was partially supported by the National Science Foundation and Department of Energy. The second author was partially supported by a National Science Foundation Postdoctoral Research Fellowship.

hierarchical controls, a high-bandwidth part of the control-structure might be turned off once the system reaches a desired basin of attraction [1].

This paper presents a fast numerical method for approximating stable and unstable manifolds of equilibrium points of a vector field. Given a smooth vector field $\boldsymbol{f}$ in $\boldsymbol{R}^n$ and a hyperbolic saddle point $\boldsymbol{y_0}$, the corresponding invariant manifolds are defined as

$$W^s(\boldsymbol{y_0}) = \left\{ \boldsymbol{y} \in \boldsymbol{R}^n \mid \lim_{t \to +\infty} \phi_f^t(\boldsymbol{y}) = \boldsymbol{y_0} \right\},$$
$$W^u(\boldsymbol{y_0}) = \left\{ \boldsymbol{y} \in \boldsymbol{R}^n \mid \lim_{t \to -\infty} \phi_f^t(\boldsymbol{y}) = \boldsymbol{y_0} \right\},$$

where ϕ_f^t is the time flow of the vector field $\boldsymbol{f}$. In the vicinity of $\boldsymbol{y_0}$, the original dynamical system

$$\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y}) \tag{1}$$

is well approximated by its linearization

$$\boldsymbol{y}' = D\boldsymbol{f}(\boldsymbol{y_0})\boldsymbol{y}. \tag{2}$$

Moreover, by the stable manifold theorem [14], the invariant manifolds of $\boldsymbol{y_0}$ are tangent to the corresponding manifolds for the linearized system (2), i.e., tangent to the respective stable ($E^s(\boldsymbol{y_0})$) and unstable ($E^u(\boldsymbol{y_0})$) eigenspaces of the matrix $D\boldsymbol{f}(\boldsymbol{y_0})$.

If the invariant manifold $W^u(\boldsymbol{y_0})$ is only one-dimensional, its approximation can be easily obtained by choosing an initial point in the unstable subspace of (2) and by integrating forward in time.[1] However, for higher-dimensional cases, the manifold consists of an infinite number of trajectories, making the task of approximating the manifold much more challenging. This problem has attracted a considerable amount of attention. In section 2 we discuss several previously available approximation methods.

We note that dynamical systems with multiple time-scales present an additional degree of complexity: for such systems, obtaining a "geometrically satisfactory" representation of the manifold is often very expensive computationally. Indeed, the most natural idea (i.e., to approximate the manifold by following a finite number of trajectories in W^u for some fixed time T) will not work very well in this case. For a simple example, consider the linear vector field

$$\boldsymbol{y}' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & -1 \end{bmatrix} \boldsymbol{y} \tag{3}$$

with a saddle point at the origin and $W^u(0)$ coinciding with the x-y plane.

Observation 1.1. In Figure 1 we show some typical trajectories in this plane and the images of a small circle around the origin under the flow ϕ_f^t. The following two well-known problems with this approach will be even more pronounced for the nonlinear case:

[1] In the rest of the paper we will mainly refer to the problem of constructing the unstable invariant manifold. The stable manifold can be constructed similarly by a time reversal (i.e., by considering the vector field $-\boldsymbol{f}$).

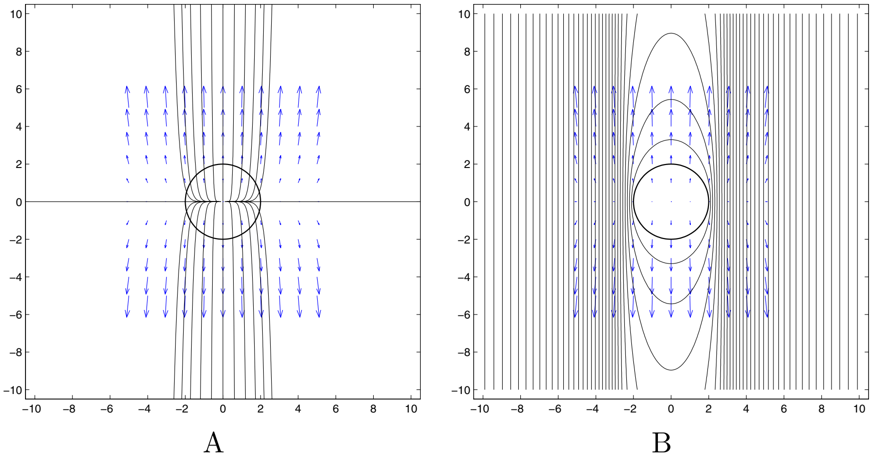


Figure 1. *A simple linear multiple-time-scale example. Trajectories are chosen to be equidistributed on a circle of radius $R_{init} = 2$.*

- If one starts with a number of points equidistributed on a small circle around $\boldsymbol{y_0}$ in $E^u(\boldsymbol{y_0})$, the speed of those points varies significantly even for relatively tame problems.
- The respective trajectories (Figure 1A) and the ϕ_f^t-images of the initial circle (Figure 1B) do not provide for a good mesh-representation of the manifold object.

The latter reflects an inherent conflict between the objectives of respecting the flow direction (which, after all, defines the manifold) and stressing the geometric properties of the manifold in a mesh-representation.

Our approach reconciles the two objectives: To approximate an invariant manifold of codimension k, we formulate a system of k quasi-linear PDEs satisfied by the manifold's local parameterization (section 3); we then solve that system locally in an Eulerian framework (section 4), thus limiting possible distortions of the mesh due to a variance of speeds for different directions inside the manifold.

The resulting discretized equations are solved very efficiently by decoupling them and ordering the computation of simplex-patches (added to the earlier computed manifold representation) to take advantage of the direction of information flow (section 6). Our algorithm can be viewed as an extension of the ordered upwind methods (OUMs) introduced in [32, 33] for static Hamilton–Jacobi PDEs. These methods solve a boundary value problem in $O(N \log N)$ operations, where N is the total number of mesh-points (section 5). The resulting triangulated mesh approximates a compact subset

$$\mathrm{W}^u_{\Sigma}(\boldsymbol{y_0}) = \{\boldsymbol{y} \in W^u(\boldsymbol{y_0}) \mid \sigma(\boldsymbol{y}) \leq \Sigma\},$$

where $\sigma(\boldsymbol{y})$ is defined as a distance-along-the-trajectory from $\boldsymbol{y}$ to $\boldsymbol{y_0}$, and Σ is a prespecified stopping criterion. For a desired mesh-scale Δ, the compactness of $\mathrm{W}^u_{\Sigma}(\boldsymbol{y_0})$ and the nondegeneracy of simplex-patches ensure that $N = O(\Delta^{-k})$.

In section 7 we use our method to construct a two-dimensional stable manifold of the origin for the Lorenz system. In section 8 we consider a dynamical system describing two pendula coupled by a torsional spring and compute two-dimensional unstable and stable manifolds for one of its saddle equilibria. Though the rate of convergence of the method is not proven, in section 9 we provide numerical evidence to confirm the first-order global accuracy. This is consistent with the local truncation error of order $O(\Delta^2)$ analytically derived in section 4.1. We conclude by discussing the limitations of our approach and possible future extensions in section 10.

2. Prior methods.

A number of previously available techniques for computing invariant manifolds all follow the same general principle, which is that an invariant k-dimensional manifold is grown as a one-parameter family $\{M_i\}$ of topological $(k-1)$-spheres, where M_0 is taken to be a small $(k-1)$-sphere around $\boldsymbol{y_0}$ in $E^u(\boldsymbol{y_0})$. However, the resulting methods are quite diverse as a result of different choices for

- the family-parameter of $\{M_i\}$ (e.g., integration time, distance along the trajectory to $\boldsymbol{y_0}$, geodesic distance to $\boldsymbol{y_0}$, etc.),
- data structures used to store the M_i's, and
- the algorithm for producing M_{i+1} given M_i.

The simplest implementation of this idea was illustrated in section 1 as follows: M_0 is approximated by a finite number of equidistant markers, the family-parameter is chosen to be the integration time, and M_{i+1} is approximated by the position of markers approximating M_i after some time Δt. As shown in Figure 1A, initially equidistributed markers quickly converge and/or drift apart due to the geometric stiffness[2] of the vector field $\boldsymbol{f}$; thus, an additional step of redistributing markers along M_{i+1} is required. Moreover, since the size of the M_i's varies, additional markers might be needed to ensure the quality of approximation (e.g., the maximum distance Δx between adjacent markers in M_{i+1}). As a result, an accurate approximation will require that small Δt be used even if marker-trajectories are computed with infinite precision.

Another problem with this approach is a highly nonuniform distance between M_i and M_{i+1} (see Figure 1B), resulting in a poor geometric approximation of the manifold even if each $(k-1)$-sphere in the family is known perfectly. A method for alleviating this difficulty was introduced by Johnson, Jolly, and Kevrekidis in [18]. Their method uses a rescaling of the vector field ("reparameterizing to integrate with respect to arclength in space-time"), which ensures the same speed along all the trajectories; i.e., the family-parameter becomes a distance-from-$\boldsymbol{y_0}$-along-the-trajectory. Unfortunately, the produced mesh still need not be the best geometric representation of the manifold since the local distance between M_i and M_{i+1} is now determined by the ratio of "M_i-normal" and "M_i-tangential" components of the rescaled vector field. We note that the computationally expensive marker-redistribution is still required at each step due to the geometric stiffness (since the rescaling of the vector field does not change the trajectories).

A method ensuring the constant distance between M_i and M_{i+1} was introduced by Guckenheimer and Worfolk in [15]. The family-parameter is chosen to be the geodesic distance from $\boldsymbol{y_0}$. The markers on M_i are moved with a unit speed for a time Δt in the direction outwards-normal to M_i within the locally determined tangent k-plane. The approximation resulting for M_{i+1} might still require marker addition/redistribution, but only due to the different size

[2] Here and throughout the paper, by *geometric stiffness* we mean the highly nonuniform rates of separation for nearby trajectories on different parts of the manifold.

of M_{i+1} and not due to the geometric stiffness. A tangent k-plane is locally determined for each marker in M_i using the adjacent markers and the direction of the vector field. Thus, this procedure becomes very sensitive wherever $\boldsymbol{f}$ is nearly tangential to M_i, leading to excessively expensive restrictions for the ratio of $\Delta t/\Delta x$.

Another method also using the geodesic distance as a family-parameter was introduced by Krauskopf and Osinga in [20, 21, 22]. For a given marker $\boldsymbol{y} \in M_i$, a locally normal $(n-k+1)$-plane $\mathcal{F}_{\boldsymbol{y}}$ is determined using the adjacent nodes in M_i. Then a shooting method is used to solve the following boundary value problem: Find a point (not necessarily a marker!) $\boldsymbol{z} \in M_i$ such that its trajectory intersects $\mathcal{F}_{\boldsymbol{y}}$ at some point $\tilde{\boldsymbol{z}}$ and $\|\tilde{\boldsymbol{z}} - \boldsymbol{y}\| = \Delta t$. A collection of $\tilde{\boldsymbol{z}}$'s is used as an approximation of M_{i+1}; as before, some new markers might be required due to the bigger size of M_{i+1}. An explicit bound on the overall computational error is available, and the quality of the resulting mesh is ensured by adding/removing the markers on M_i, depending on the manifold's local geometry [21]. We note that the above procedure is robust even if the vector field $\boldsymbol{f}$ is locally tangential to M_i, but the shooting method becomes much more computationally expensive in that case. In addition, solving the boundary value problem for each marker becomes even more challenging for $k > 2$ because the search space for $\boldsymbol{z}$ is no longer one-dimensional.

Remark 2.1. All of the above methods are explicit in the sense that only the representation of M_i is used to produce M_{i+1} and the order of computation of the markers on M_{i+1} is unimportant. Thus, these methods' computational complexity is generally proportional to the total number (across all of the M_i's) of used mesh-points N. However, N will depend not only on the required accuracy in manifold-approximation but also on the choice of family-parameter. Moreover, the proportionality constants involved can be quite large, depending on k (e.g., for the marker-redistribution) and on the orientation of $\boldsymbol{f}$ relative to the M_i's.

Several other numerical techniques are not based on growing a family of M_i's.

A method introduced by Doedel [11] uses a single computed trajectory in $W^u(\boldsymbol{y_0})$ as an input for the boundary value solver of AUTO [10] to perform continuation in the ray-angle parameter. The manifold is approximated between $M_{init} = M_0$ and M_{final} by a sequence of trajectories $\{\boldsymbol{z^j}\}$ such that $\boldsymbol{z^j}(0) \in M_0$, $\boldsymbol{z^j}(\tau_j) \in M_{final}$, and $\|\boldsymbol{z^j}(\tau_j) - \boldsymbol{z^{j+1}}(\tau_{j+1})\| = \Delta$. Starting with the initial trajectory $\boldsymbol{z^0}$, the collocation methods are repeatedly used to produce $\boldsymbol{z^{j+1}}$ based on $\boldsymbol{z^j}$. If P_j is a hyperplane transversal to $\boldsymbol{f}$ at $\boldsymbol{z^j}(\tau_j)$, then $\boldsymbol{z^{j+1}}$ is sought with one end point on M_0 and the other lying on P_j distance Δ away from $\boldsymbol{z^j}(\tau_j)$. The resulting sequence $\{\boldsymbol{z^j}\}$ is well spaced near M_{final} but may not be uniformly spaced near M_0.

A new method by Henderson [16] is based on integrating an individual trajectory together with a second-order approximation to the manifold along that trajectory. The surface is constructed as a collection of k-dimensional strips centered at such trajectories; the use of these (nonintersecting) strips provides uniform bounds on the spacing of the trajectories. The implementation heavily relies on the efficient data structures developed earlier for approximating implicitly defined manifolds [17]. This method is the first to directly model the curvature information in the direction transversal to the trajectories.

An algorithm introduced by Dellnitz and Hohmann in [6, 7] uses subdivision and cell-mapping-continuation techniques to produce an n-dimensional covering of the k-dimensional unstable manifold. A simplified version of this algorithm can be summarized as follows. The computational domain Q is subdivided into a number of small, nonintersecting n-dimensional "boxes." The initial covering is determined as a collection of those boxes covering $W^u_{loc}(\boldsymbol{y_0})$—a small neighborhood of $\boldsymbol{y_0}$ in $W^u(\boldsymbol{y_0})$. The iteratively repeated continuation stage adds new boxes to the collection if they are intersected by a $\phi_f^{\Delta t}$-image of some box(es) already in the collection. The process stops when no more boxes within Q can be added. The use of efficient (hierarchical) data structures allows storing only those boxes actually needed for the covering. The covering's growth reflects the anisotropy due to multiple time-scales present in the system; i.e., the more strongly unstable directions are covered first. The accuracy of the approximation depends upon the size of boxes in the resulting covering and upon the level of refinement of the initial covering (the relative sizes of the boxes compared to $W^u_{loc}(\boldsymbol{y_0})$); hence the algorithm can be quite memory intensive and may converge relatively slowly, especially for $k = n-1$. The efficiency of this algorithm also strongly depends on the contraction transversal to the manifold: weaker contraction will require a much finer initial covering—otherwise, the cell mapping will produce a coarse n-dimensional covering of W^u.

We note that the method in [6, 7] is currently the only one implemented for $k > 2$. Several other methods briefly described above were formulated for the general case, but, to the best of our knowledge, the current implementations rely on $k = 2$.

3. PDE approach to manifold-approximation. In contrast to the methods discussed in the previous section, we compute an invariant manifold as a collection of adjacent k-dimensional simplex-patches. The $(k-1)$-dimensional boundary of the current collection is used to attach new tentative simplexes, whose exact position in $\boldsymbol{R}^n$ is computed using a PDE for the local parameterization of that manifold.

We begin by considering a relatively simple case of a two-dimensional manifold in $\boldsymbol{R}^3$. If $(x_1, x_2, g(x_1, x_2)) = (\boldsymbol{x}, g(\boldsymbol{x})) = \boldsymbol{y}$ is a local parameterization of an invariant manifold, then the vector field evaluated on it should be tangential to the graph of $g(x_1, x_2)$, i.e.,

$$\boldsymbol{f}\,(x_1, x_2, g(x_1, x_2)) \begin{bmatrix} \frac{\partial}{\partial x_1} g(x_1, x_2) \\ \frac{\partial}{\partial x_2} g(x_1, x_2) \\ -1 \end{bmatrix} = 0, \tag{4}$$

should hold wherever this parameterization is valid. Our general method can be outlined as follows:

- The above first-order quasi-linear PDE can be solved to "continue" the manifold since the boundary condition for g is specified on the last previously computed manifold "boundary."
- The initial "boundary" is approximated by a discretized small circle around $\boldsymbol{y_0}$ in $E^u(\boldsymbol{y_0})$.
- Once a new triangle-patch of the manifold is computed and *Accepted*, the computational "boundary" ($AcceptedFront$) is modified to include it, new tentative (or $Considered$) patches are added to the computational domain, and the PDE is solved on them using the new (local) coordinates.

This process is discussed in detail in section 6 and illustrated in section 7. Here, we simply note that, unlike a general quasi-linear PDE, equation (4) always has a smooth solution as long as the chosen parameterization remains valid. Thus, switching to local coordinates when solving the PDE allows us to avoid checking the continued validity of the parameterization.

The above derivation can be repeated to obtain a single PDE defining an invariant manifold of codimension 1 in $\boldsymbol{R}^n$: the number of equations corresponds to the number of linearly independent vectors orthogonal to the manifold's tangent space, i.e., to the manifold's codimension.

In this spirit, we now consider the general problem of constructing a k-dimensional invariant manifold of a vector field $\boldsymbol{f} : \boldsymbol{R}^n \to \boldsymbol{R}^n$. Switching to a suitable coordinate system, we assume that the manifold's local parameterization is $(x_1, \dots, x_k, g_1(x_1, \dots, x_k), \dots, g_{n-k}(x_1, \dots, x_k)) = (\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{y} \in \boldsymbol{R}^n$, where $\boldsymbol{x} \in \boldsymbol{R}^k$ and $\boldsymbol{g} : \boldsymbol{R}^k \mapsto \boldsymbol{R}^{n-k}$. As in the codimension 1 case, the PDE is derived from the condition that the vector field evaluated on the manifold should lie in its tangent space. Therefore, for every $j \in \{1, \dots, (n-k)\}$,

$$\sum_{i=1}^{k} \frac{\partial g_j}{\partial x_i}(x_1, \dots, x_k) f_i(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})) = f_{j+k}(\boldsymbol{x}, \boldsymbol{g}(\boldsymbol{x})) \tag{5}$$

should locally hold as long as the above parameterization is valid. This coupled system of $(n-k)$ quasi-linear PDEs can again be used to "continue" the manifold since the boundary conditions for the g_j's are specified on the last-previously-computed-manifold-"boundary." In this case, the initial "boundary" is approximated by a discretized $(k-1)$-sphere around $\boldsymbol{y_0}$ in $E^u(\boldsymbol{y_0})$ and the manifold grows as new k-dimensional simplexes are *Accepted*. The construction of manifolds of codimension 2 is illustrated in section 8.

Remark 3.1 (a historical note).The PDE approach for characterizing invariant surfaces goes back to at least the 1960s. In particular, the existence and smoothness of solutions for equations equivalent to (5) are the subjects of Sacker's analytical perturbation theory [30, 25]. Previous numerical techniques based on this formulation included time-marching finite difference schemes in "special coordinates" [9], iterative methods [8] based on a discrete version of Fenichel's graph transform [13], collocation methods [12], and spectral methods [24]. However, all this work was done for invariant tori computations, resulting in the following two very important distinctions from our method:

1. These prior methods assume the existence of a coordinate system in which the invariant torus is indeed globally a graph of the function. Such a coordinate system may be defined explicitly [9] or implicitly [8]. In the latter case it can be defined using normal/tangent bundles of a (previously constructed) invariant torus of a slightly perturbed vector field. This implies availability of a global mesh on which the PDE can be solved.

   For invariant manifolds of hyperbolic equilibria, such a mesh is not available a priori and has to be constructed in the process of "growing" the manifold (section 6.5).
2. For the invariant tori computations, the solution function $\boldsymbol{g}$ has periodic boundary conditions; hence, the discretized equations are inherently coupled and have to be solved simultaneously.

   For approximation of $W^u(\boldsymbol{y_0})$, all characteristics of the PDE start at the initial boundary (chosen in $E^u(\boldsymbol{y_0})$) and run "outward." Knowledge of the direction of information flow can be used to decouple the discretized system, resulting in a much faster computational method (section 5).

4. Eulerian discretization. Not surprisingly, the characteristics of (4) are exactly the (projections of the) trajectories of the original vector field. Thus, any attempt to solve it in the Lagrangian framework (i.e., by the method of characteristics or ray shooting) would bring us back to all the problems discussed in section 1. On the other hand, it was demonstrated in [33] that the discretized (semi-Lagrangian and Eulerian) equations resulting from certain nonlinear first-order PDEs can be solved very efficiently. This was our motivation for locally recasting this problem in a fully Eulerian framework.

For a two-dimensional invariant manifold in $\boldsymbol{R}^3$ (as formulated in (4)), let $G(x_1, x_2)$ be a piecewise-linear numerical approximation of the solution $g(x_1, x_2)$. Consider a simplex $\boldsymbol{y}\boldsymbol{y^1}\boldsymbol{y^2}$, where $\boldsymbol{y^i} = (x_1^i, x_2^i, G(x_1^i, x_2^i)) = (\boldsymbol{x^i}, G(\boldsymbol{x^i}))$ and $\boldsymbol{y} = (x_1, x_2, G(x_1, x_2)) = (\boldsymbol{x}, G(\boldsymbol{x}))$. We assume that the vertices $\boldsymbol{y^1}$ and $\boldsymbol{y^2}$ are two adjacent mesh-points on the $AcceptedFront$ (the discretization of the current manifold "boundary"). Thus, $G(\boldsymbol{x^i})$'s are known and can be used in computing $G(\boldsymbol{x})$. Define the unit vectors $\boldsymbol{P_i} = (\boldsymbol{x} - \boldsymbol{x^i})/\|\boldsymbol{x} - \boldsymbol{x^i}\|$ and let P be a matrix with $\boldsymbol{P_i}$'s as its rows. This square matrix is invertible since $\boldsymbol{x}$ is chosen some distance away from the $AcceptedFront$. We note that a directional derivative for G in the direction $\boldsymbol{P_i}$ can be computed as

$$v_i(\boldsymbol{x}) = \frac{G(\boldsymbol{x}) - G(\boldsymbol{x^i})}{\|\boldsymbol{x} - \boldsymbol{x^i}\|}. \tag{6}$$

Therefore, if $\boldsymbol{v}$ is a column vector of v_i's, then $\nabla g(\boldsymbol{x}) \approx \nabla G(\boldsymbol{x}) = P^{-1}\boldsymbol{v}$, yielding the discretized version of (4):

$$\left[P^{-1}\boldsymbol{v}(\boldsymbol{x})\right]_1 f_1(\boldsymbol{x}, G(\boldsymbol{x})) \;+\; \left[P^{-1}\boldsymbol{v}(\boldsymbol{x})\right]_2 f_2(\boldsymbol{x}, G(\boldsymbol{x})) \;=\; f_3(\boldsymbol{x}, G(\boldsymbol{x})). \tag{7}$$

This nonlinear equation can be solved for $G(\boldsymbol{x})$ by the Newton–Raphson method or any other robust zero-solver. In addition, it has an especially simple geometric interpretation if the local coordinates are chosen so that $G(\boldsymbol{x^1}) = G(\boldsymbol{x^2}) = 0$. Setting $\hat{\boldsymbol{y}} = (\boldsymbol{x}, 0)$, we reduce the problem to finding the correct "tilt" for a simplex $\boldsymbol{y}\boldsymbol{y^1}\boldsymbol{y^2}$. If $\boldsymbol{u}$ is a unit vector normal to $\hat{\boldsymbol{y}}\boldsymbol{y^1}\boldsymbol{y^2}$, then solving (7) is equivalent to finding a number $\alpha \in \boldsymbol{R}$ such that $\boldsymbol{f}(\hat{\boldsymbol{y}} + \alpha\boldsymbol{u})$ lies in the plane defined by $\boldsymbol{y^1}$, $\boldsymbol{y^2}$, and $\boldsymbol{y} = \hat{\boldsymbol{y}} + \alpha\boldsymbol{u}$. (See Figure 2.)

This geometric interpretation can be extended to the general case[3] of a k-dimensional invariant manifold in $\boldsymbol{R}^n$. In this case, the $AcceptedFront$ is a $(k-1)$-dimensional mesh discretizing the currently computed manifold "boundary" and we consider a k-dimensional simplex $\boldsymbol{y}\boldsymbol{y^1}\ldots\boldsymbol{y^k}$, where $\boldsymbol{y^1}, \ldots, \boldsymbol{y^k} \in \boldsymbol{R}^n$ form a $(k-1)$-dimensional simplex in $AcceptedFront$ and $\boldsymbol{y}$ is a $Considered$ point near it. A local parameterization $\boldsymbol{g}(\boldsymbol{x})$ satisfying system (5) is numerically approximated by $\boldsymbol{G}(\boldsymbol{x})$; i.e., we assume $\boldsymbol{y^i} = (\boldsymbol{x^i}, \boldsymbol{G}(\boldsymbol{x^i}))$ and $\boldsymbol{y} = (\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}))$, where $\boldsymbol{x}, \boldsymbol{x^i} \in \boldsymbol{R}^k$ and $\boldsymbol{G} : \boldsymbol{R}^k \to \boldsymbol{R}^{n-k}$. We choose the parameterization so that $\boldsymbol{G}(\boldsymbol{x^i}) = \boldsymbol{0}$ and let $\hat{\boldsymbol{y}} = (\boldsymbol{x}, \boldsymbol{0})$. Let $\{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^{(n-k)}\}$ form an orthonormal basis for the orthogonal complement of the k-plane $\hat{\boldsymbol{y}}\boldsymbol{y^1}\ldots\boldsymbol{y^k}$. Then the task of linearly approximating system (5) is equivalent to finding the real numbers $\alpha_1, \ldots, \alpha_{(n-k)}$ such that, for $\boldsymbol{y} = \hat{\boldsymbol{y}} + \sum_{i=1}^{n-k} \alpha_i \boldsymbol{u}^i$, the vector $\boldsymbol{f}(\boldsymbol{y})$ lies in the k-plane defined by $\boldsymbol{y}, \boldsymbol{y^1}, \ldots, \boldsymbol{y^k}$.

[3] Of course, the explicit discretization formula is also available. We omit it here for the sake of brevity and notational clarity.

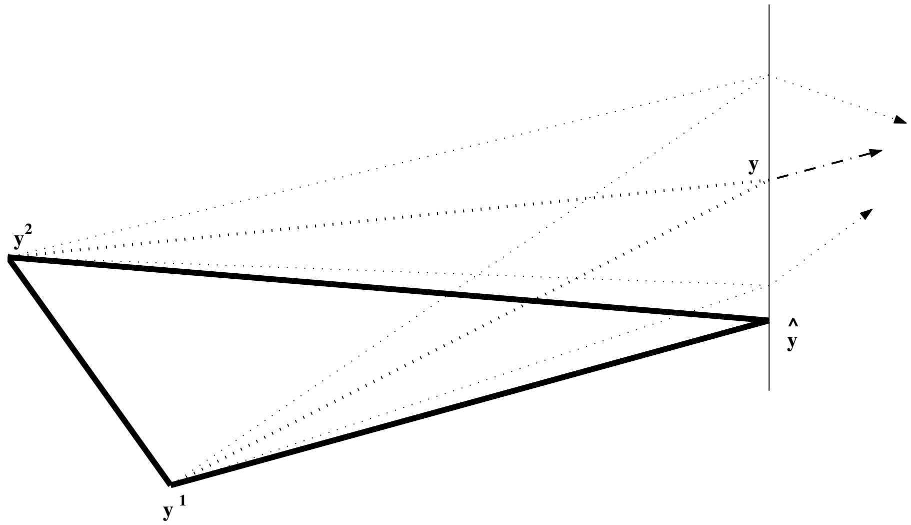


Figure 2. *Geometric interpretation of (7). The one-dimensional search-space corresponds to the manifold's codimension.*

The described discretization procedure is similar in spirit to an *implicit Euler method* for solving initial value problems since $\boldsymbol{y^i}$'s are assumed to be known and the vector field is computed at a yet-to-be-determined point $\boldsymbol{y}$.

4.1. Local truncation error and upwinding condition. Let L be the Lipschitz constant of $\boldsymbol{f}$ and let ν be the upper bound of $\|\nabla g\|$ on a Δ-neighborhood of $\boldsymbol{x}$. For $k = 2$, suppose that $\boldsymbol{y^1} = (\boldsymbol{x^1}, G(\boldsymbol{x^1}))$ and $\boldsymbol{y^2} = (\boldsymbol{x^2}, G(\boldsymbol{x^2}))$ lie on the manifold and that $\boldsymbol{y} = (\boldsymbol{x}, G(\boldsymbol{x}))$ solves (7). This means that $\boldsymbol{f}(\boldsymbol{y})$ lies in the plane of $\boldsymbol{y}\boldsymbol{y^1}\boldsymbol{y^2}$ and can be expressed as a linear combination

$$\boldsymbol{f}(\boldsymbol{y}) = \beta_1(\boldsymbol{y} - \boldsymbol{y^1}) + \beta_2(\boldsymbol{y} - \boldsymbol{y^2}), \tag{8}$$

where the real coefficients β_1 and β_2 satisfy the equation

$$P^T \left[\begin{array}{c} \beta_1\|\boldsymbol{x} - \boldsymbol{x^1}\| \\ \beta_2\|\boldsymbol{x} - \boldsymbol{x^2}\| \end{array} \right] = \left[\begin{array}{c} f_1(\boldsymbol{y}) \\ f_2(\boldsymbol{y}) \end{array} \right]. \tag{9}$$

If $\boldsymbol{f}(\boldsymbol{y})$ is not parallel to $\boldsymbol{y^1}\boldsymbol{y^2}$, then $\beta_1 + \beta_2 \neq 0$, and a linear approximation of $\boldsymbol{y}$'s trajectory (i.e., the straight line through $\boldsymbol{y}$ in the direction $\boldsymbol{f}(\boldsymbol{y})$) will intersect the line $\boldsymbol{y^1}\boldsymbol{y^2}$ at the point $\tilde{\boldsymbol{y}} = (\beta_1\boldsymbol{y^1} + \beta_2\boldsymbol{y^2})/(\beta_1 + \beta_2)$. We note that

$$\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| = \frac{\|\boldsymbol{f}(\boldsymbol{y})\|}{|\beta_1 + \beta_2|}. \tag{10}$$

Since, away from equilibria, $\|\boldsymbol{f}(\boldsymbol{y})\|$ is bounded from below, (10) implies $|\beta_1 + \beta_2|^{-1} = O(\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|)$.

Using the above notation, we can rewrite (7) in the form

$$G(\boldsymbol{x}) = \frac{\beta_1}{\beta_1 + \beta_2} G(\boldsymbol{x^1}) + \frac{\beta_2}{\beta_1 + \beta_2} G(\boldsymbol{x^2}) + \frac{f_3(\boldsymbol{x}, G(\boldsymbol{x}))}{\beta_1 + \beta_2}. \tag{11}$$

For $\tilde{\boldsymbol{x}} = (\beta_1 \boldsymbol{x^1} + \beta_2 \boldsymbol{x^2})/(\beta_1 + \beta_2)$ we can now express

$$\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2 \;=\; \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2 + \left(\frac{f_3(\boldsymbol{x}, G(\boldsymbol{x}))}{\beta_1 + \beta_2}\right)^2 \;=\; \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2 + \left(\frac{f_3(\boldsymbol{y})\,\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|}{\|\boldsymbol{f}(\boldsymbol{y})\|}\right)^2.$$

Therefore,

$$\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2 \;=\; \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2 \frac{\|\boldsymbol{f}(\boldsymbol{y})\|^2}{\|\boldsymbol{f}(\boldsymbol{y})\|^2 - (f_3(\boldsymbol{y}))^2}.$$

Thus, $O(\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|) = O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|)$, provided $|f_3| \ll \|\boldsymbol{f}\|$. This condition can be satisfied by a suitable choice of the coordinate system, e.g., if the point $\boldsymbol{x}$ is chosen so that $f_3(\tilde{\boldsymbol{x}}) = 0$. In that case, the "correction term" $f_3(\boldsymbol{y})/(\beta_1 + \beta_2)$ in formula (11) is of the order $O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2)$. For the solution $g(\boldsymbol{x})$ of PDE (4),

$$\begin{aligned} g(\boldsymbol{x}) &= g(\tilde{\boldsymbol{x}}) + \nabla g(\boldsymbol{x}) \cdot (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2) \\ &= g(\tilde{\boldsymbol{x}}) + \nabla g(\boldsymbol{x}) \cdot \frac{\beta_1(\boldsymbol{x} - \boldsymbol{x^1}) + \beta_2(\boldsymbol{x} - \boldsymbol{x^2})}{\beta_1 + \beta_2} + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2). \end{aligned}$$

Combining the above with (8), we obtain

$$g(\boldsymbol{x}) = g(\tilde{\boldsymbol{x}}) + \frac{f_1(\boldsymbol{x}, G(\boldsymbol{x}))\frac{\partial}{\partial x_1} g(\boldsymbol{x}) + f_2(\boldsymbol{x}, G(\boldsymbol{x}))\frac{\partial}{\partial x_2} g(\boldsymbol{x})}{\beta_1 + \beta_2} + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2)$$

and

$$\begin{aligned} |g(\boldsymbol{x}) - G(\boldsymbol{x})| &\le \left| g(\tilde{\boldsymbol{x}}) + \frac{f_1(\boldsymbol{x}, G(\boldsymbol{x}))\frac{\partial}{\partial x_1} g(\boldsymbol{x}) + f_2(\boldsymbol{x}, G(\boldsymbol{x}))\frac{\partial}{\partial x_2} g(\boldsymbol{x})}{\beta_1 + \beta_2} - G(\boldsymbol{x}) \right| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2) \\ &\le \left| g(\tilde{\boldsymbol{x}}) + \frac{f_1(\boldsymbol{x}, g(\boldsymbol{x}))\frac{\partial}{\partial x_1} g(\boldsymbol{x}) + f_2(\boldsymbol{x}, g(\boldsymbol{x}))\frac{\partial}{\partial x_2} g(\boldsymbol{x})}{\beta_1 + \beta_2} - G(\boldsymbol{x}) \right| \\ &\quad + \frac{2L\nu}{|\beta_1 + \beta_2|} |g(\boldsymbol{x}) - G(\boldsymbol{x})| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2). \end{aligned}$$

Since g solves the PDE, we obtain

$$\begin{aligned} |g(\boldsymbol{x}) - G(\boldsymbol{x})| &\le \left| g(\tilde{\boldsymbol{x}}) + \frac{f_3(\boldsymbol{x}, g(\boldsymbol{x}))}{\beta_1 + \beta_2} - G(\boldsymbol{x}) \right| + \frac{2L\nu}{|\beta_1 + \beta_2|} |g(\boldsymbol{x}) - G(\boldsymbol{x})| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2) \\ &\le \left| g(\tilde{\boldsymbol{x}}) + \frac{f_3(\boldsymbol{x}, G(\boldsymbol{x}))}{\beta_1 + \beta_2} - G(\boldsymbol{x}) \right| + \frac{L(2\nu + 1)}{|\beta_1 + \beta_2|} |g(\boldsymbol{x}) - G(\boldsymbol{x})| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2). \end{aligned}$$

Setting $C = L(2\nu + 1)/\|\boldsymbol{f}(\boldsymbol{y})\|$ and recalling (10), (11) gives

$$\begin{aligned} (1 - C\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|)\; |g(\boldsymbol{x}) - G(\boldsymbol{x})| &\le \left| g(\tilde{\boldsymbol{x}}) + \frac{f_3(\boldsymbol{x}, G(\boldsymbol{x}))}{\beta_1 + \beta_2} - G(\boldsymbol{x}) \right| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2) \\ &= \left| g(\tilde{\boldsymbol{x}}) - \frac{\beta_1}{\beta_1 + \beta_2} G(\boldsymbol{x^1}) - \frac{\beta_2}{\beta_1 + \beta_2} G(\boldsymbol{x^2}) \right| + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2). \end{aligned}$$

Let $\boldsymbol{x}^m = (\boldsymbol{x}^{\boldsymbol{1}} + \boldsymbol{x}^{\boldsymbol{2}})/2$. Since $\tilde{\boldsymbol{x}}$ is on the line $\boldsymbol{x}^{\boldsymbol{1}}\boldsymbol{x}^{\boldsymbol{2}}$ and it was assumed that $G(\boldsymbol{x}^{\boldsymbol{i}}) = g(\boldsymbol{x}^{\boldsymbol{i}})$, the linear approximation yields

$$g(\tilde{\boldsymbol{x}}) = (\beta_1 G(\boldsymbol{x}^{\boldsymbol{1}}) + \beta_2 G(\boldsymbol{x}^{\boldsymbol{2}}))/(\beta_1 + \beta_2) + O(\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^m\|^2) + O(\|\boldsymbol{x}^{\boldsymbol{1}} - \boldsymbol{x}^{\boldsymbol{2}}\|^2).$$

Thus,

$$|g(\boldsymbol{x}) - G(\boldsymbol{x})| \leq \frac{O\left(\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^m\|^2\right) + O\left(\|\boldsymbol{x}^{\boldsymbol{1}} - \boldsymbol{x}^{\boldsymbol{2}}\|^2\right) + O(\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2)}{(1 - C\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|)}. \tag{12}$$

If $\tilde{\boldsymbol{x}}$ lies in between $\boldsymbol{x}^{\boldsymbol{1}}$ and $\boldsymbol{x}^{\boldsymbol{2}}$, and if the triangle $\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{1}}\boldsymbol{x}^{\boldsymbol{2}}$ has sides of length $\leq \Delta$, then formula (12) yields a local truncation error of order $O(\Delta^2)$. Correspondingly, we expect the global approximation error of order $O(\Delta)$ for the entire mesh. The rigorous analysis of the global error is outside the scope of this paper, but in section 9 we provide numerical evidence to confirm the first-order accuracy.

The requirement that $\tilde{\boldsymbol{x}}$ should lie in between $\boldsymbol{x}^{\boldsymbol{1}}$ and $\boldsymbol{x}^{\boldsymbol{2}}$ ensures that interpolation (rather than extrapolation) is used for $G(\tilde{\boldsymbol{x}})$. Moreover, it corresponds to the fundamental stability condition for solving first-order PDEs: the mathematical domain of dependence should be included in the numerical domain of dependence. For our problem this means that $G(\boldsymbol{x})$ should be computed using the correct triangle—the triangle through which the corresponding (approximate) trajectory runs. Thus, having computed $\boldsymbol{y} = (\boldsymbol{x}, G(\boldsymbol{x}))$ by (7) using two adjacent mesh-points $\boldsymbol{y}^{\boldsymbol{i}}$ and $\boldsymbol{y}^{\boldsymbol{j}}$, we need to verify the following additional *upwinding condition*: the linear approximation to $\boldsymbol{y}$'s trajectory should intersect the line $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$ at the point $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{x}}, G(\tilde{\boldsymbol{x}}))$ lying between $\boldsymbol{y}^{\boldsymbol{i}}$ and $\boldsymbol{y}^{\boldsymbol{j}}$ (see Figure 3) or, equivalently, $\boldsymbol{f}(\boldsymbol{y})$ should point from the newly computed simplex $\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$. If the upwinding criterion is satisfied (i.e., $\beta_1, \beta_2 \geq 0$), formula (10) provides the length of the linear approximation of $\boldsymbol{y}$'s trajectory inside $\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$. Both the upwinding condition and formula (10) can be similarly extended for $k > 2$.

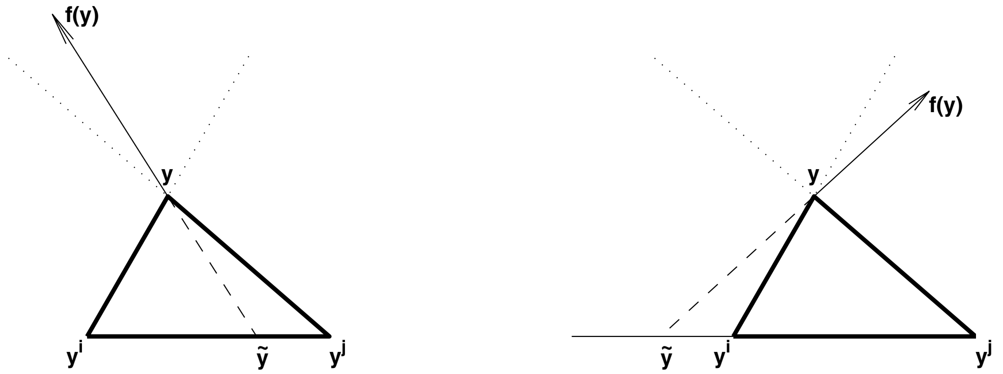


Figure 3. *Examples of acceptable (left) and unacceptable (right) approximations of $\boldsymbol{f}(\boldsymbol{y})$. The range of upwindings directions is shown by dotted lines; the local linear approximation to the trajectory is shown by dashed line; $\tilde{\boldsymbol{y}}$ is its intersection with the line $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$. In the second case the upwinding criterion is not satisfied and the update for $\boldsymbol{y}$ should be computed using another segment of AcceptedFront.*

5. Ordered upwind methods. OUMs were originally introduced by Sethian and Vladimirsky to solve a class of problems in anisotropic control theory and anisotropic front propagation described by static Hamilton–Jacobi–Bellman PDEs [32, 33]. A finite-difference discretization of a nonlinear boundary value problem normally leads to a system of N nonlinear coupled discretized equations, where N is the total number of mesh-points in the computational domain. The solution to that system is usually obtained iteratively, while each iteration involves recomputing the values at all of the mesh-points. Such iterative schemes can be quite slow even in conjunction with Gauss–Seidel relaxation techniques. OUMs provide an alternative by using the partial information about the direction of information flow to essentially decouple the system and to solve the equations one by one. Several extensions of these methods were introduced for hybrid control problems [34] and for phase-space multiple-arrivals computations [31].

The decoupling introduced in [32] hinges on the notion of "optimality" and the variational properties of the PDEs arising in the control-theoretic context. This formulation was heavily used in proving convergence to the viscosity solution of the Hamilton–Jacobi–Bellman PDE [33]. However, the more general idea behind the methods was to allow *space-marching* for the boundary value problems—not unlike explicit forward-time marching for initial-boundary value problems. In essence, the solution can be "marched" (on the mesh) from the boundary using the characteristic information, and a new (smaller) boundary value problem can be posed using the newly computed "boundary"—the current divide between the already-computed ($Accepted$) and not-yet-touched (Far) mesh-points. The mesh discretization of that "new boundary" is referred to as $AcceptedFront$; the not-yet-$Accepted$ mesh-points, which are adjacent to the $AcceptedFront$, are designated $Considered$. A tentative value can be computed for each $Considered$ mesh-point $\boldsymbol{x}$ under the assumption that its characteristic intersects the $AcceptedFront$ in some vicinity of that mesh-point (designated $NF(\boldsymbol{x})$). All $Considered$ points are sorted based on the $SortValue$ (usually defined as the time-to-travel to $\boldsymbol{x}$ from the boundary along its characteristic). A typical step of the algorithm consists of choosing the $Considered$ $\bar{\boldsymbol{x}}$ with the smallest $SortValue$ and making it $Accepted$. This operation modifies the $AcceptedFront$ ($\bar{\boldsymbol{x}}$ in; other mesh-points possibly out), and causes a possible recomputation of all the not-yet-$Accepted$ mesh-points near $\bar{\boldsymbol{x}}$.

This "space-marching" is based on the principle of "local" solution reconstruction from characteristics and on some notion of an entropy-like condition (i.e., no characteristics emerging from shocks). Both of these are applicable for a much wider class of first-order PDEs. In [31] OUMs were successfully used to treat the linear Liouville PDE. The applicability of OUMs to general quasi-linear first-order PDEs is still an open question [36]. However, the particular computational problem considered in this paper has an additional simplifying property: the PDEs (4) and (5) are solved only locally, and hence the solution remains smooth at every point. On the other hand, unlike in the previous OUMs, the mesh is not known in advance and is built in the process of computation. In adding a tentative simplex-patch (with a $Considered$ vertex) we attempt to provide for "good" geometric properties of the mesh (e.g., simplex aspect ratio) and to ensure that the parameterization is locally well-conditioned (e.g., $\|\nabla g\|$ should be small on that simplex). The vector field near $AcceptedFront$ determines the order in which the correct "tilts" for tentative simplex-patches are computed and the $Considered$ mesh-points are $Accepted$. This ordering has the effect of reducing the approximation error (a mesh-point $\boldsymbol{y}$ first computed from a relatively far part of $NF(\boldsymbol{y})$ is likely to be recomputed before it gets *Accepted*). Below we outline the general structure of the algorithm and provide a detailed description of individual components in section 6. As in the original OUMs, the computational complexity of the algorithm is $O(N \log N)$, where the $(\log N)$ factor results from the necessity of maintaining a sorted list of $Considered$ mesh-points.

Ordered upwind method for building invariant manifolds.

1. Use the linearization $(E^u(\boldsymbol{y_0}))$ to initialize $AcceptedFront$ and one "layer" of $Considered$s.
2. Evaluate the tentative coordinates for $Considered$s.
3. Find the $Considered$ mesh-point $\bar{\boldsymbol{y}}$ which is the closest (in the sense of trajectory distances) to $AcceptedFront$.
4. Move $\bar{\boldsymbol{y}}$ to *Accepted* and update the $AcceptedFront$.
5. Remove/add the $Considered$ mesh-points to reflect changes to $AcceptedFront$.
6. Recompute the coordinates for all the $Considered$ $\boldsymbol{y}$ such that $\bar{\boldsymbol{y}} \in NF(\boldsymbol{y})$.
7. If $Considered$ is not empty (and "stopping criteria" are not met) then go to 3.

6. Implementation details. The current implementation is specifically geared toward two-dimensional manifolds in $\boldsymbol{R}^n$, even though a generalization of most of the following is straightforward (except for section 6.3 and parts of section 6.5 which explicitly rely on $k = 2$). Our goal is to construct a simplicial complex approximating the manifold with the preferred triangle side of length Δ (but definitely less than 2Δ). Throughout this section we call a mesh-triangle s *adjacent* to $\boldsymbol{y}$ if $\boldsymbol{y}$ is one of the vertices of s; we also refer to mesh-points $\boldsymbol{y^i}$ and $\boldsymbol{y^j}$ as adjacent (or connected) if both of them are vertices of the same triangle.

6.1. Initialization. The algorithm is initialized using the linearization of the vector field. $E^u(\boldsymbol{y_0})$ is determined as a span of eigenvectors of $D\boldsymbol{f}(\boldsymbol{y_0})$ corresponding to the eigenvalues with positive real part.

Any simple closed curve around $\boldsymbol{y_0}$ in $E^u(\boldsymbol{y_0})$ can be used as an initial boundary I, provided that curve is transversal to the linearized vector field. In the examples considered in the following sections, I was chosen to be a circle of radius R_{init} centered at $\boldsymbol{y_0}$. More generally, the transversality condition can always be satisfied by choosing an ellipse corresponding to the relevant eigenvectors.

The initial boundary I is then approximated using N_0 *Accepted* mesh-points so that the distance between all adjacent mesh-points $\boldsymbol{y^1}$ and $\boldsymbol{y^2}$ is at most Δ. For each such adjacent pair, the segment $\boldsymbol{y^1 y^2}$ is placed onto the $AcceptedFront$ and an equilateral triangle is constructed with the vertices at $\boldsymbol{y^1}$, $\boldsymbol{y^2}$, and $\hat{\boldsymbol{y}}$, where the new $Considered$ mesh-point $\hat{\boldsymbol{y}}$ lies in $E^u(\boldsymbol{y_0})$ and $\|\boldsymbol{y_0} - \hat{\boldsymbol{y}}\| = R_{init} + \Delta\sqrt{3}/2$. (See Figure 4.)

6.2. Computing coordinates for $Considered$s. Once the correct "tilt" is computed, each $Considered$ mesh point $\boldsymbol{y}$ is a vertex of a triangle with the other vertices $\boldsymbol{y^1}$, $\boldsymbol{y^2}$ on the $AcceptedFront$. Initially, however, that triangle is built to be locally tangential to the manifold; i.e., only $\hat{\boldsymbol{y}}$ is known at first instead of $\boldsymbol{y}$ (see Figure 2). If $\boldsymbol{y^1}$ and $\boldsymbol{y^2}$ are on I, then $\hat{\boldsymbol{y}}$ is chosen in $E^u(\boldsymbol{y_0})$, as described above; otherwise, the position of $\hat{\boldsymbol{y}}$ is selected in the plane of the previously *Accepted* triangle adjacent to $\boldsymbol{y^1}$ and $\boldsymbol{y^2}$ (as described in section 6.5). Given $\hat{\boldsymbol{y}}$, the discretized version of the PDE(s) is solved to obtain the "normal component(s)" of $\boldsymbol{y}$.

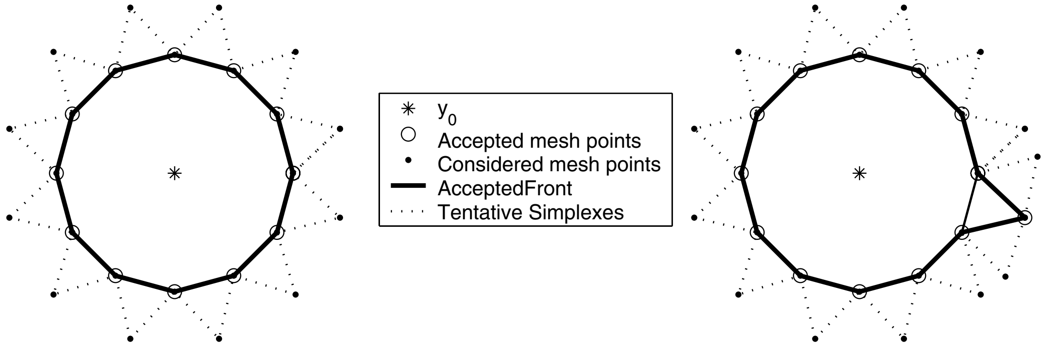


Figure 4. *Initialization of AcceptedFront in $E^u(\boldsymbol{y_0})$ (left) and changes to AcceptedFront once the first of Considered points is Accepted (right). The newly added/accepted triangle generally does not lie in $E^u(\boldsymbol{y_0})$; once Accepted, it defines the plane in which the new tentative triangles are initially chosen.*

As described in section 4, the newly computed $\boldsymbol{y}$ is a valid *Considered* point if it satisfies the "upwinding condition." If that condition is not satisfied, we recompute $\boldsymbol{y}$ using other segments on *AcceptedFront* near $\hat{\boldsymbol{y}}$. In general, given two adjacent mesh-points $\boldsymbol{y^i}$ and $\boldsymbol{y^j}$ on *AcceptedFront*, we can form a "virtual simplex" $\hat{\boldsymbol{y}}\boldsymbol{y^i}\boldsymbol{y^j}$ which is then used to compute the value for $\boldsymbol{y}$ even if it is not directly adjacent to $\boldsymbol{y^i}$ or $\boldsymbol{y^j}$. A *NearFront* $NF(\hat{\boldsymbol{y}})$ is defined as a collection of segments on *AcceptedFront* within the distance R_{NF} from $\hat{\boldsymbol{y}}$.

$NF(\hat{\boldsymbol{y}})$ is used to restrict the set of virtual simplexes, which will be potentially checked to find the one satisfying the upwinding criterion—the vector $\boldsymbol{f}(\boldsymbol{y})$ should be pointing out of the simplex used to compute $\boldsymbol{y}$ (Figure 5 illustrates this for a simplified case $\hat{\boldsymbol{y}} = \boldsymbol{y}$).

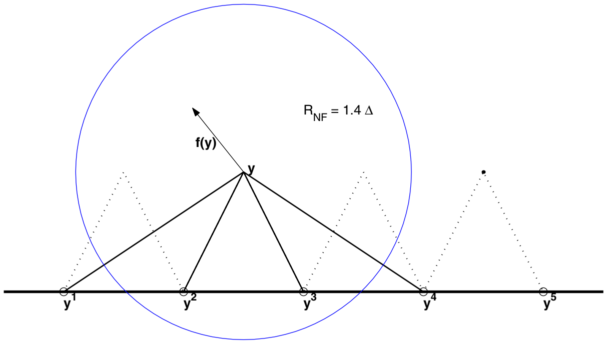


Figure 5. *Use of $NF(\hat{\boldsymbol{y}})$ to build virtual simplexes for evaluating $\boldsymbol{y}$. The first evaluation is performed using $\boldsymbol{y^2}$ and $\boldsymbol{y^3}$ as the Accepted mesh-points adjacent to $\hat{\boldsymbol{y}}$; the resulting approximation for $\boldsymbol{f}(\boldsymbol{y})$ shows that the upwinding condition is not satisfied and that a virtual simplex using $\boldsymbol{y^3}\boldsymbol{y^4}$ should be considered next.*

Remark 6.1. Generally, one needs to select the value for R_{NF} based on the behavior of the vector field near *AcceptedFront*—if the locally tangential components dominate the locally normal components, it will not be possible to satisfy the upwinding criterion unless R_{NF} is sufficiently large. (Note that this will also increase the local truncation error of section 4.1 by a factor of R_{NF}.)

On the other hand, this truly becomes a problem only if this local-tangentiality holds everywhere along the $AcceptedFront$: the fact that $Considered$ mesh-points are ordered based on $SortValue$ allows for a subsequent recomputation of coordinates of $\boldsymbol{y}$ once the position of $AcceptedFront$ changes. (See Figure 6.)

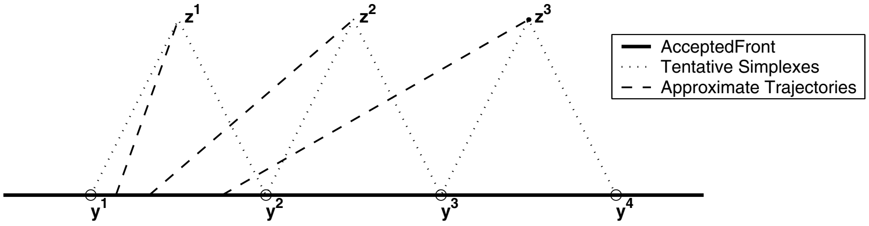


Figure 6. *Ordering acceptance of Considered based on SortValue decreases the computational stencil; this reduces both the minimum sufficient R_{NF} and the local truncation error. Even though all three $\boldsymbol{z^i}$'s can be computed using $\boldsymbol{y^1y^2}$, this is not really necessary. Once $\boldsymbol{z^1}$ becomes Accepted, both $\boldsymbol{z^2}$ and $\boldsymbol{z^3}$ can be computed from $\boldsymbol{z^1y^2}$; once $\boldsymbol{z^2}$ becomes Accepted, $\boldsymbol{z^3}$ can be recomputed using $\boldsymbol{z^2y^3}$. Thus, in this example, valid coordinate updates will be eventually computed even if $R_{NF} = \Delta$.*

6.3. Relaxing the upwinding condition.

The $AcceptedFront$ is a one-dimensional object, and the very first evaluation of $\boldsymbol{y}$'s coordinates will indicate the correct search-direction to satisfy the upwinding condition (Figure 5).

Unfortunately, $\boldsymbol{f}(\boldsymbol{y})$ gives only an approximation of the direction of information flow since the "normal" components (i.e., $\boldsymbol{y} - \hat{\boldsymbol{y}}$) are computed numerically from the first-order accurate discretization of the PDE. As a result, when the adjacent segments in $NF(\hat{\boldsymbol{y}})$ do not lie on the same line, it is possible that the upwinding condition will not be satisfied by any virtual simplex. (See Figure 7).

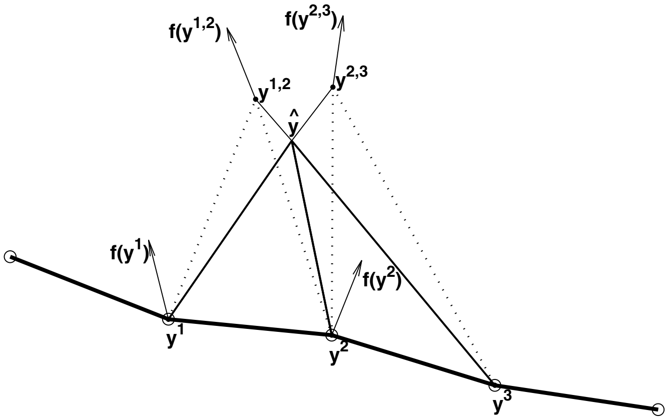


Figure 7. *Deadlock situation: according to $\boldsymbol{f}(\boldsymbol{y^{1,2}})$, $\boldsymbol{y}$'s trajectory should intersect $\boldsymbol{y^2y^3}$; according to $\boldsymbol{f}(\boldsymbol{y^{2,3}})$, the trajectory should intersect $\boldsymbol{y^1y^2}$. Thus, neither virtual simplex satisfies the strict upwinding condition and the relaxation of the criteria is needed. The directions of $\boldsymbol{f}(\boldsymbol{y^1})$ and $\boldsymbol{f}(\boldsymbol{y^2})$ show that the update $\boldsymbol{y^{1,2}}$ satisfies the relaxed criterion.*

Remark 6.2. Thus, we use the following relaxation of the upwinding criterion: Let $\hat{\boldsymbol{y}}$ be a $Considered$ point and let $\boldsymbol{y}^{\boldsymbol{i},\boldsymbol{j}}$ be its new coordinates computed from a virtual simplex $s = \hat{\boldsymbol{y}}\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$. The *relaxed upwinding condition* is satisfied if there exists a point $\tilde{\boldsymbol{y}}$ on an $AcceptedFront$ segment $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$ such that the projection of $\boldsymbol{f}(\tilde{\boldsymbol{y}})$ onto s is collinear with $\tilde{\boldsymbol{y}}\hat{\boldsymbol{y}}$. In practice, an alternative version of this condition is easier to verify: it suffices to check that the projections of $\boldsymbol{f}(\boldsymbol{y}^{\boldsymbol{i}})$ and $\boldsymbol{f}(\boldsymbol{y}^{\boldsymbol{j}})$ onto the plane of s lie outside of that simplex.

Remark 6.3. As formulated in section 4.1, the upwinding criterion is implicit: it cannot be verified until the tentative value $\boldsymbol{y}^{\boldsymbol{i},\boldsymbol{j}}$ is computed. In contrast, the above *relaxed upwinding criterion* is explicit in nature since it concerns only the directions of the vector field on $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$.

We note that, even when an explicit (relaxed) upwinding criterion is used, the "tilt" $(\boldsymbol{y} - \hat{\boldsymbol{y}})$ is still computed from the implicit formula (7). Moreover, in such cases we still have $\|\tilde{\boldsymbol{y}} - \boldsymbol{y}\| = O(R_{NF}\Delta)$ and the local truncation error derived in section 4.1 is still valid.

The above reasoning clearly uses the fact that $k = 2$. In general, the $AcceptedFront$ will be a $(k-1)$-dimensional object and both the search in $NF(\hat{\boldsymbol{y}})$ and the upwinding-relaxation procedures will have to be more complicated.

In our implementation, the relaxed upwinding is only used to deal with the deadlocks at $Considered$ points tagged as $Lagging$ (immediately adjacent to more than two segments of $AcceptedFront$ yet possessing no valid update). In all other cases, relaxation is postponed since subsequent modifications to $AcceptedFront$ may allow for the strict upwinding condition to be satisfied. As a result, the relaxation is applied very infrequently (e.g., at 24 mesh-points out of 77,500 in the example considered in section 8).

6.4. Computing the $SortValue$ and sorting $Considered$s. Our decoupling orders the acceptance of $Considered$ points based on their "distance-along-the-trajectory-to-$\boldsymbol{y_0}$." That distance $\sigma(\boldsymbol{y})$ can be estimated as a sum of the "distance-along-the-trajectory-to-$AcceptedFront$" and $\sigma(\tilde{\boldsymbol{y}})$, where $\tilde{\boldsymbol{y}}$ is the intersection of $\boldsymbol{y}$'s trajectory with a segment $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$ on the $AcceptedFront$ (see Figure 3).

Once the new coordinates for a $Considered$ point $\boldsymbol{y}$ are computed and the upwinding criterion is satisfied, we obtain a linear approximation to $\boldsymbol{y}$'s trajectory and can estimate its $SortValue$: $\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|$ can be computed by formula (10) and $\sigma(\tilde{\boldsymbol{y}})$ can be approximated by linearly interpolating $\sigma(\cdot)$ on $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$. Since $\boldsymbol{y}$ is computed from $\boldsymbol{y}^{\boldsymbol{i}}\boldsymbol{y}^{\boldsymbol{j}}$, we recall that $\boldsymbol{f}(\boldsymbol{y}) = \beta_1(\boldsymbol{y} - \boldsymbol{y}^{\boldsymbol{i}}) + \beta_2(\boldsymbol{y} - \boldsymbol{y}^{\boldsymbol{j}})$ for some $\beta_1, \beta_2 \geq 0$ and

$$(13) \qquad SortValue(\boldsymbol{y}) = \sigma(\boldsymbol{y}) \;\approx\; \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| + \sigma(\tilde{\boldsymbol{y}}) \;\approx\; \frac{\|\boldsymbol{f}(\boldsymbol{y})\| + \beta_1\sigma(\boldsymbol{y}^{\boldsymbol{i}}) + \beta_2\sigma(\boldsymbol{y}^{\boldsymbol{j}})}{\beta_1 + \beta_2}.$$

If the (relaxed) upwinding criterion for $\boldsymbol{y}$ cannot be satisfied by any segment in $NF(\boldsymbol{y})$, then we leave $\boldsymbol{y} = \hat{\boldsymbol{y}}$ and assume $SortValue(\boldsymbol{y}) = +\infty$. Such a $Considered$ point will never get $Accepted$ unless the upwinding criterion is later satisfied in the subsequent recomputations (triggered by changes to $AcceptedFront$ near $\boldsymbol{y}$).

We use a heap-sort data structure to maintain the sorting of the $Considered$ points based on their $SortValue$s. As a result, selecting $\bar{\boldsymbol{y}}$ (stage 3 of the algorithm) can be performed in $O(1)$ operations, but every time a $Considered$ point's $SortValue$ changes, its position in the heap-sort should change as well. This resorting can be performed in $O(\log K)$ operations, where $K < N$ is the current number of $Considered$ mesh-points.

Remark 6.4. As noted in section 5, the general OUMs require using $T(\boldsymbol{y})$, the time-to-travel-along-the-characteristic, as a $SortValue$. However, this choice is dictated by the necessity of building the "correct" global (weak) solution to the PDE—any other ordering will risk running through the shocks and/or violating the entropy conditions [36]. In our current work, the characteristics are the trajectories of a smooth vector field and the solution is computed only locally; hence shocks cannot occur if Δ is sufficiently small. This smoothness of the solution enables us to use different sorting criteria, including $\sigma(\boldsymbol{y})$ (as above), $d(\boldsymbol{y})$ (the geodesic distance-to-$\boldsymbol{y_0}$), and the geodesic (or along-the-trajectory) distance to $AcceptedFront$. Our particular choice ($SortValue = \sigma(\boldsymbol{y})$) is a result of an empirical trade-off: it requires a lesser R_{NF} than ($SortValue = d(\boldsymbol{y})$) and generally decreases the length of $AcceptedFront$ as compared to ($SortValue = T(\boldsymbol{y})$).

6.5. Changing $AcceptedFront$ and extending the mesh.

Every $Considered$ point $\boldsymbol{y}$ is a vertex of at least one *tentative triangle* $\boldsymbol{y}\boldsymbol{y^1}\boldsymbol{y^2}$, where $\boldsymbol{y^1}$ and $\boldsymbol{y^2}$ are adjacent mesh-points on the $AcceptedFront$. As a $Considered$ mesh-point $\bar{\boldsymbol{y}}$ becomes $Accepted$, this tentative triangle becomes *fully accepted* and the $AcceptedFront$ has to be modified accordingly. (Note that the newly added triangle will always use the segment $\boldsymbol{y^1}\boldsymbol{y^2}$ adjacent to $\bar{\boldsymbol{y}}$—even if that point was computed using some other segment $\boldsymbol{y^i}\boldsymbol{y^j}$ in $NF(\bar{\boldsymbol{y}})$.) The changes to $AcceptedFront$ proceed in two stages as follows:

1. Removal from $AcceptedFront$ of each segment $\boldsymbol{y^i}\boldsymbol{y^j}$ shared by two fully accepted triangles (or used by just one such triangle if both $\boldsymbol{y^i}$ and $\boldsymbol{y^j}$ are on the initial boundary I).
2. Adding to $AcceptedFront$ segments $\bar{\boldsymbol{y}}\boldsymbol{y^j}$, for all $AcceptedFront$ mesh-points $\boldsymbol{y^j}$ adjacent to $\bar{\boldsymbol{y}}$.

Once the $AcceptedFront$ has been modified, it may be necessary to extend the mesh near $\bar{\boldsymbol{y}}$. The existing mesh includes $Accepted$ points and a narrow band of $Considered$ points near $AcceptedFront$. If two mesh-points $\boldsymbol{y^k}, \boldsymbol{y^l} \notin I$ are adjacent, then we will refer to $\boldsymbol{y^k}\boldsymbol{y^l}$ as a *perimeter segment* if that segment is used as an edge by a unique triangle. The $AcceptedFront$ plays the role of an approximate boundary for (locally) solving the PDE. Correspondingly, if $\boldsymbol{y}$ is $Accepted$ but not on $AcceptedFront$, then there should be no perimeter segments adjacent to $\boldsymbol{y}$. Moreover, if $\boldsymbol{y^k}\boldsymbol{y^l}$ is on $AcceptedFront$, then it should not be a perimeter segment either. Yet if one of these points was just $Accepted$, the segment may be on the perimeter of the existing mesh and some local mesh-building is required to create the second triangle adjacent to $\boldsymbol{y^k}\boldsymbol{y^l}$. The following heuristic algorithm is similar to the "advancing front mesh generation" method described in [27].

Local mesh-extension algorithm. Let $L = \|\boldsymbol{y^k} - \boldsymbol{y^l}\|$ be the length of a perimeter segment $\boldsymbol{y^k}\boldsymbol{y^l}$. Let γ be the smallest *outer* angle formed by this segment, i.e., $\gamma = \min(\angle \boldsymbol{y^j}\boldsymbol{y^k}\boldsymbol{y^l}, \angle \boldsymbol{y^k}\boldsymbol{y^l}\boldsymbol{y^m})$, where $\boldsymbol{y^j}\boldsymbol{y^k}$ and $\boldsymbol{y^l}\boldsymbol{y^m}$ are perimeter segments adjacent to $\boldsymbol{y^k}\boldsymbol{y^l}$. Also, since $\boldsymbol{y^k}, \boldsymbol{y^l} \notin I$, there already exists a unique fully accepted triangle s_{kl} with these two vertices.

A second triangle adjacent to $\boldsymbol{y^k}\boldsymbol{y^l}$ is created by one of the following three procedures:

1. If $\gamma \geq \pi$, then we introduce a new mesh-point $\hat{\boldsymbol{y}}$ (in the plane defined by s_{kl}) to form

an isosceles triangle with $\boldsymbol{y^k y^l}$ as its base and sides $\boldsymbol{y^k \hat{y}}$, $\boldsymbol{y^l \hat{y}}$ of length L_1, where

$$L_1 = \begin{cases} 2\,L & \text{if } \; L \leq \frac{\Delta}{2}; \\ \Delta & \text{if } \; \frac{\Delta}{2} < L \leq \frac{\Delta}{0.55}; \\ 0.55\,L & \text{if } \; \frac{\Delta}{0.55} < L. \end{cases}$$

(See Figure 8A.) The constants used above are purely heuristic and are intended to balance our preference for nearly regular triangles against the desired mesh-scale Δ; see [27] for further details.

2. If $\angle \boldsymbol{y^j y^k y^l} = \gamma$ is acute, then a triangle $\boldsymbol{y^j y^k y^l}$ is added to the mesh. If $\|\boldsymbol{y^j y^l}\| > 2\Delta$, then that triangle is split in two by adding a new *Considered* mesh-point $\boldsymbol{\hat{y}}$. (See Figure 8B.)
3. If $\gamma \in [\frac{\pi}{2}, \pi)$, then we compute R_{jkl} and R_{klm}, the radii of circles passing through the respective triples of points. (The radius is assumed $+\infty$ if the corresponding outer angle is $\geq \pi$.) Without loss of generality, assume that $R_{jkl} \geq R_{klm}$. If $R_{klm} < L_1$, then a triangle $\boldsymbol{y^k y^l y^m}$ is added to the mesh without adding any new mesh-points. (See Figure 8C.) Otherwise, we create a new isosceles triangle $\boldsymbol{\hat{y} y^k y^l}$, as described above (see procedure 1).

We note that if a new mesh-point is created, then the choice of $\boldsymbol{\hat{y}}$ merely fixes the local coordinate system in which the "normal" components of $\boldsymbol{y}$ are next computed from $NF(\boldsymbol{y})$ (stage 6 of the algorithm) as described in section 6.2.

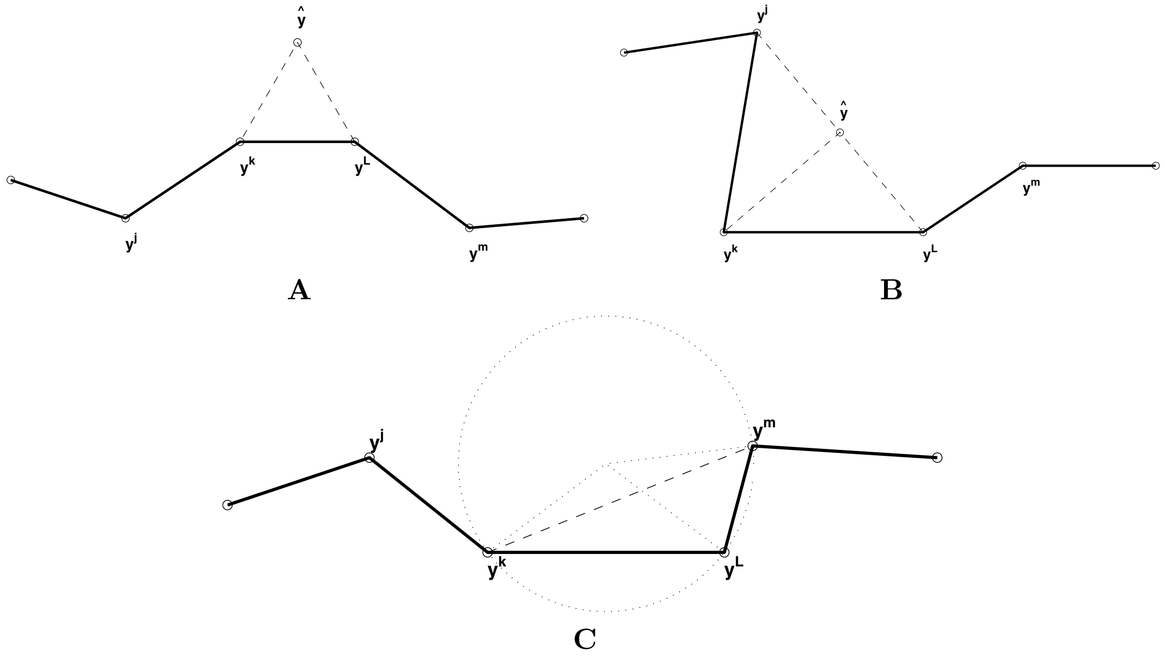


Figure 8. *Local mesh generation: three different procedures for extending the mesh at $\boldsymbol{y^k y^l}$. New segments are shown by dashed lines. If created, the new mesh-point is labeled $\boldsymbol{\hat{y}}$. In all examples, the mesh is assumed preexistent below the polygonal perimeter.*

Remark 6.5. The local mesh-generation step above provides no provable guarantee for the quality of the resulting triangles (aspect ratio, minimum angle, etc.). Nevertheless, we note that

1. in practice, the generated mesh is fairly well behaved (e.g., in the example considered in section 7, the minimum angle present in the mesh is $> 18°$ and the vast majority of triangles have minimum angles $> 30°$.)
2. the aspect ratio of the constructed triangles does not directly affect the quality of manifold-approximation. Every *Considered* mesh-point $\boldsymbol{y}$ is computed using all the segments in $NF(\boldsymbol{y})$, not only the immediately adjacent triangles.
3. additional postprocessing procedures (*diagonal swapping* and *mesh smoothing*) can be used to improve the aspect ratios once the manifold is constructed. See [27] for further details.
4. our mesh-generation method requires $O(1)$ operations for each *Accepted* point. More sophisticated (and more computationally expensive) mesh-generation procedures can be used to obtain triangles with guaranteed minimum angles. For example, mesh-quality guarantees can be obtained for Delaunay triangulation methods [5] and for hybrid advancing-front/Delaunay triangulation methods [29].
5. our method clearly uses the fact that $k = 2$. For the general case $k > 2$, the mesh extension step would require building a (local) simplicial complex in the manifold tangent space compatible with the current polytope boundary (i.e., $AcceptedFront$). This construction has to be performed only locally and no mesh-quality (aspect ratio) guarantees are required. Thus, any hypersurface meshing method can be used (e.g., [4, 17]).

6.6. Stopping criteria.

The stopping criterion used in our implementation is the unavailability of any *Considered* points with $\sigma(\boldsymbol{y}) \leq \Sigma$, where Σ is a prespecified (max-distance-along-the-trajectory) parameter. Using the heap-sort data structure, the criterion can be checked by a single comparison: the algorithm stops as soon as $SortValue(\bar{\boldsymbol{y}}) > \Sigma$, where $\bar{\boldsymbol{y}}$ is the current first element on the heap.

We note that other stopping criteria (Euclidean or geodesic distance, time-along-trajectory, total number of simplexes, etc.) can be used independently of the chosen $SortValue$.

6.7. Algorithm features and possible optimizations.

The algorithm we have described has the worst-case computational complexity of $O(R_{NF} N \log K)$, where N is the total number of mesh-points, R_{NF} provides the maximum number of recomputations for a $Considered$ point, and K is the maximum number of mesh-points marked $Considered$ at the same time (typically, $\approx \sqrt{N}$). This is different from the $O(N)$ complexity of the explicit methods of section 2. Nevertheless, this additional cost is justified since it results in a reduction of discretization errors; see, for example, Figure 6 and the discussion of local truncation errors in section 4.1. In addition, the overall computational efficiency of our method is much better since the "constant coefficients terms" in the computational cost are largely dependent on the geometry of the manifold rather than on the geometric stiffness of the vector field (see Remark 2.1).

Remark 6.6. For $k > 2$, a generalization of the algorithm described above will have the same asymptotic complexity of $O(R_{NF} N \log K)$. However, a (constant factor) increase in cost appears for $k < n-1$. In that case, a system of $(n-k)$ PDEs has to be solved to update each $Considered$ point (section 3). The geometric argument in section 4 shows that solving the discretization of that system requires an $(n-k)$-dimensional Newton–Raphson method. This contrasts our method with the approaches introduced in [18, 21, 7], for which the computational cost of updating a single marker increases with the manifold dimension.

Remark 6.7. Given our method of construction, the manifold-approximation always contains the (approximate) trajectory of each already $Accepted$ mesh-point. (This stems from the upwinding criteria and is independent of our choices of $SortValue$ and/or stopping criteria.) That property is similarly possessed by the methods introduced in [18] and [6], but not by those in [15] and [20].

Remark 6.8. Our algorithm may terminate before Σ is actually reached: given the particular choices for the method-parameters (desired mesh-scale Δ, initialization radius R_{init}, and $NearFront$ radius R_{NF}), it might be impossible to obtain an accurate (upwinding-condition-satisfying) update for any of the current $Considered$ points. Such $Considered$ points will have $SortValue = +\infty$ and, for the sake of efficiency, will not be placed onto the heap-sort. We note that such "early termination" can be easily determined in $O(K)$ operations by checking if $\sigma(\boldsymbol{y^j}) + \Delta << \Sigma$ for any $\boldsymbol{y^j}$ on the final $AcceptedFront$. For the examples in this paper, the suitable parameter values were obtained empirically. A better implementation would address this adaptively: R_{NF} can be increased and/or Δ can be decreased (by refining the current $AcceptedFront$) whenever a possibility of early termination is detected. In addition, a valuable extension would be to vary these parameters automatically based on the detected information about the manifold geometry (e.g., curvature), the vector field's tangency to $AcceptedFront$, and on the current error estimate. We note that such adaptive versions are already available for some of the prior methods mentioned in section 2 (see, e.g., [21]).

7. Example: The Lorenz system.

We consider the classical example of the Lorenz system [23]:

$$\begin{aligned} x' &= \varsigma(y-x); \\ y' &= \rho x - y - xz; \\ z' &= -\beta z + xy; \end{aligned} \tag{14}$$

with the canonical parameter values $\varsigma = 10$, $\beta = \frac{8}{3}$, and $\rho = 28$.

In this case, the system has three fixed points: the origin and $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The eigenvalues for the Jacobian at the origin are $\lambda \approx -22.8$, -2.67, 11.8; thus, the origin has a two-dimensional stable manifold. The ratio of the eigenvalues suggests (at least locally) the geometric stiffness similar to that encountered in (3).

In addition, the stable manifold of the origin has complicated geometry: it spirals into the famous "butterfly-like" chaotic attractor and also twists around the z-axis; see Figure 9. As a result, it became a de facto standard for testing methods for the invariant manifold-approximation (e.g., compare with [22, 15, 7, 16]).

We initialize the $AcceptedFront$ by subdividing the circle of radius $R_{init} = 2$ around the origin in $E^s(0)$ into $N_0 = 21$ segments (i.e., $\Delta = 0.6$). We start by placing 21 "hanging-simplexes" into the list of $Considered$s and proceed as described in section 5. The calculation

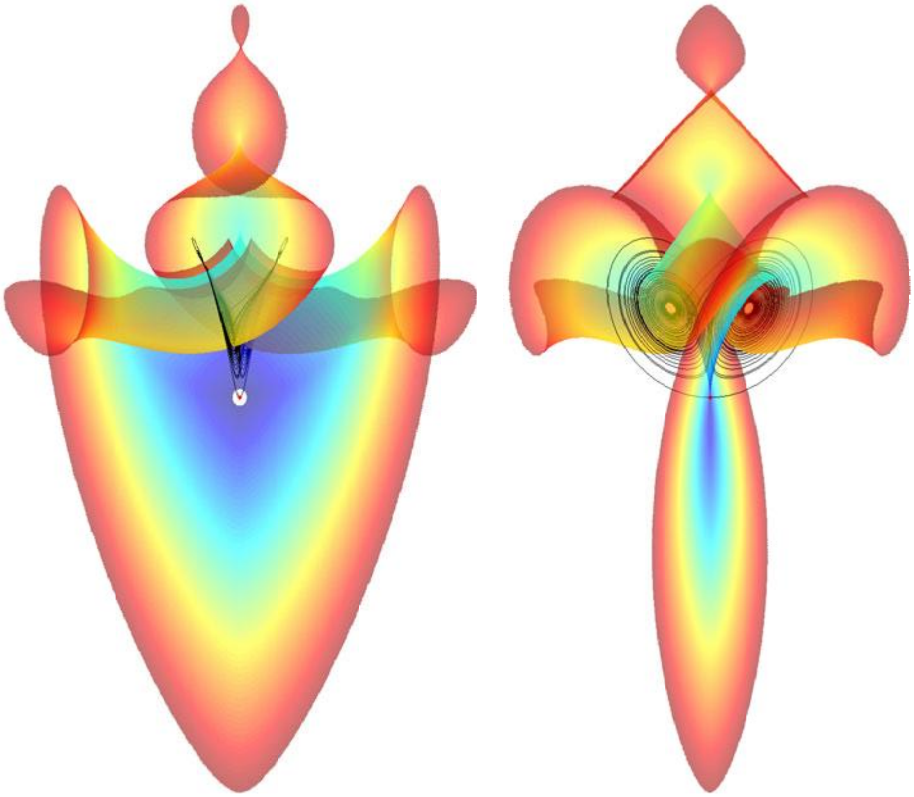


Figure 9. *The invariant manifolds of the origin (two views; rotated around z-axis). The stable manifold (displayed semitransparent) is computed up to $\Sigma = 120$ and the color indicates the trajectory-arc-length σ. The unstable manifold (in black) is computed by integrating initial conditions in $E^u(0)$ forward in time.*

stops once we *Accept* everything with the trajectory-arc-length less than the specified Σ. For the computation in Figure 10, R_{NF} was set to 4Δ; the resulting mesh contained 116,082 mesh-points and 230,011 simplexes.

Based on the visual and numerical evidence, the produced triangulated surfaces seem to converge to $\mathrm{W}^s_\Sigma(0)$ as the accuracy parameters (R_{init} and Δ) tend to zero. Aside from comparing Figures 9 and 10 with those in [22], etc., we also note the indirect evidence of two sample trajectories appearing to lie on the manifold in Figure 10. These trajectories are obtained by integrating backward in time the initial conditions $\pm\varepsilon\boldsymbol{e}$, where ε is small relative to R_{init} and $\boldsymbol{e}$ is a unit eigenvector corresponding to the eigenvalue $\lambda \approx -22.8$. Animated movies visualizing the growth and structure of this manifold are available at

http://www.math.cornell.edu/~vlad/manifold_movies/lorenz.html.

Based on a $O(\Delta^2)$ local truncation error of the discretized equation (7), we expect the first-order convergence of the approximation to $W^s_\Sigma(0)$. Unfortunately, there is no known closed-form parameterization for this manifold, which makes computing global approximation errors difficult. Even computing the distance between two such triangulated surfaces (obtained for different Δ's) is not a trivial task. We rely on the examples of section 9 to numerically test the order of convergence of our method.

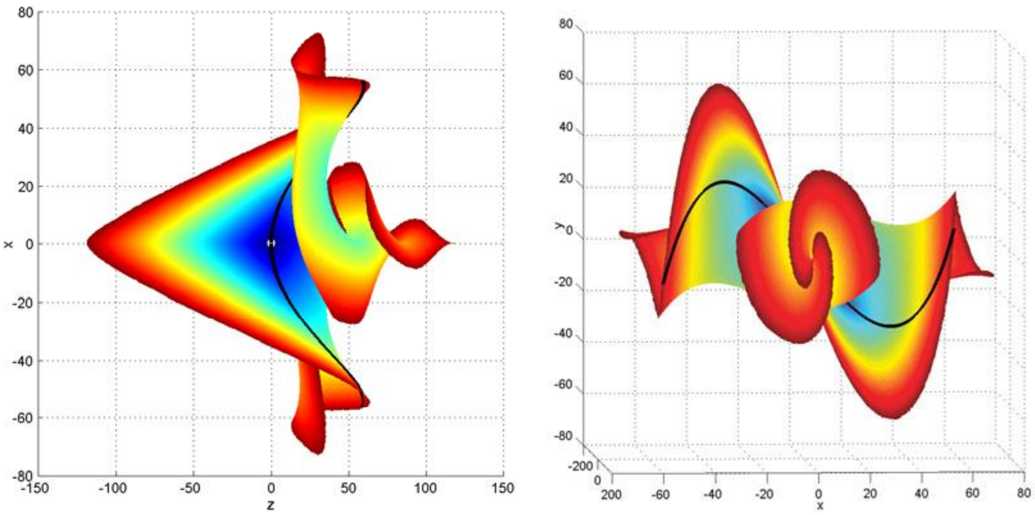


Figure 10. *Stable manifold of the origin computed up to the trajectory-arc-length $\Sigma = 120$. The color indicates σ. Two sample trajectories are shown for verification purposes.*

8. Example: Pendula coupled by torsion.

To demonstrate the applicability of our method to constructing invariant manifolds of higher codimension, we consider here a test problem of two simple pendula coupled by a torsional spring:

$$\begin{aligned} \psi_1''(t) &= -\sin(\psi_1(t)) + \varepsilon(\psi_2(t) - \psi_1(t)), \\ \psi_2''(t) &= -\sin(\psi_2(t)) + \varepsilon(\psi_1(t) - \psi_2(t)). \end{aligned} \tag{15}$$

This problem is discussed in detail in [3]; here, we reproduce only some basic properties of the system.

In (15), ψ_i is the angular position of the ith pendulum, the full state of the system can be recorded as $(\psi_1, \psi_2, \psi_1', \psi_2')$, and the full phase-space is therefore four-dimensional. As written above, the system is conservative, with the total energy given by

$$E = \frac{(\psi_1')^2}{2} + \frac{(\psi_2')^2}{2} - \cos(\psi_1) - \cos(\psi_2) + \frac{\varepsilon\,(\psi_1 - \psi_2)^2}{2}. \tag{16}$$

The constant ε corresponds to a scaled Hooke's law coefficient and we are interested in investigating the system for $\varepsilon \ll 1$ (e.g., $\varepsilon = 0.01, 0.05$). We would like to construct the invariant manifolds of the saddle point at $\boldsymbol{z^0} = (\pi, \pi, 0, 0)$ (i.e., both pendula standing upright with zero angular velocity). The eigenvalues of the Jacobian matrix are $\pm\sqrt{1 - 2\varepsilon}$ and ± 1; thus, both stable and unstable manifolds are two-dimensional and there are no multiple time-scales in the linearized system near $\boldsymbol{z^0}$. However, the energy level $E = 2$ corresponding to this saddle is singular: it contains both stable and unstable manifolds of all the equilibria of the form $\boldsymbol{z^m} = ((2m+1)\pi, (2m+1)\pi, 0, 0)$. At the same time, for $i \neq j$, this energy level *does not* contain any points of the form $((2i+1)\pi, (2j+1)\pi, \cdot, \cdot)$ —because of the torsional spring, the potential energy at those points is higher than $E(\boldsymbol{z^0})$.

Figure 11 shows the projections of the entire energy level and of sample trajectories in $W^u(\boldsymbol{z^0})$ into the configuration plane (ψ_1, ψ_2) for different values of ε. The configuration space has a periodic structure (the behavior at (ψ_1, ψ_2) is the same as at $(\psi_1+2\pi m, \psi_2+2\pi m)$). The uncoupled system (for $\varepsilon = 0$) is doubly periodic and, as a result, for small ε the configuration plane clearly has a cellular structure. The *cells* are the squares whose vertices are at the points $((2m+1)\pi, (2n+1)\pi)$ and whose boundaries correspond to the state where one of the pendula is upright and the dynamics is especially sensitive.

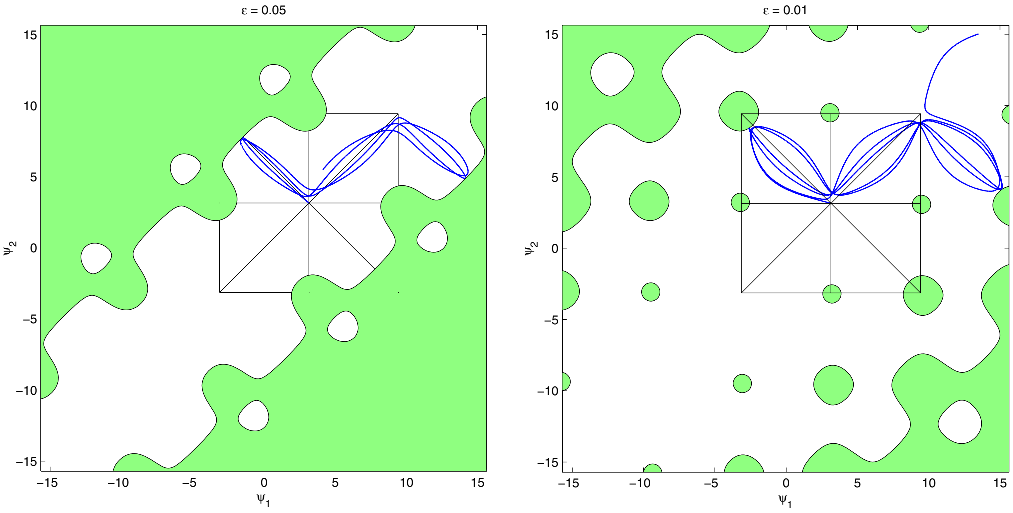


Figure 11. *Projection into the configuration space: the energy level and a single (typical) orbit in* $W^u(\boldsymbol{z^0})$ *for* $\varepsilon = 0.05$ *and* $\varepsilon = 0.01$*. The light green regions are unattainable due to the energy conservation.*

Several types of connecting orbits inside the energy level can be determined and are very useful in assessing the accuracy of the invariant manifold computations. For example, there are heteroclinic trajectories connecting each $\boldsymbol{z^m}$ with $\boldsymbol{z^{m\pm1}}$ (corresponding to the pendula moving in unison) and homoclinic orbits lying along the "antidiagonals" $\psi_1 + \psi_2 = (4m+2)\pi$ (corresponding to the pendula departing from z^m in opposite directions and moving in symmetry until the spring pulls them back).

Here, we present several views of $W^u(\boldsymbol{z^0})$ for $\varepsilon = 0.01$. Animated movies visualizing the growth and structure of the manifold for $\varepsilon = 0.1$ are available at

http://www.math.cornell.edu/~vlad/manifold_movies/pendula.html.

In our computation, we initialized the *AcceptedFront* on a circle of radius $R_{init} = 0.5$ in $E^u(\boldsymbol{z^0})$, used $\Delta = 0.1$, and computed the manifold up to $\sigma_{\max} = 15.5$. We note several algorithmic differences from the previous example. An energy conserving method could be built to essentially reduce the search-space to a single dimension (i.e., except at $\boldsymbol{z^m}$, the energy level is three-dimensional and the invariant manifolds have codimension 1 inside it). Instead, we have chosen to implement the method in the full four-dimensional phase space, thus illustrating the construction of a manifold of codimension 2. A system of two quasi-linear PDEs is solved to obtain each *Considered* point $\boldsymbol{y}$ (see (5)); the solution to the discretized nonlinear system is obtained by a standard two-dimensional Newton–Raphson method (e.g., see [28]). The resulting triangulated mesh approximates $W^u(\boldsymbol{z^0})$. However, once the *AcceptedFront* is sufficiently close to $\boldsymbol{z^{\pm1}}$, the numerical losses in energy result in "retracting" along the

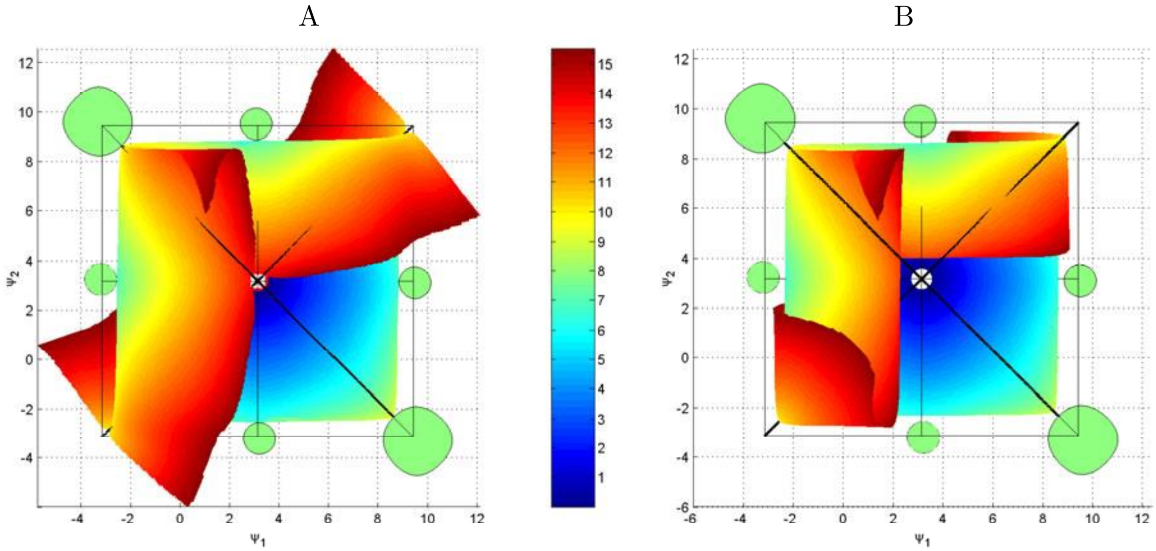


Figure 12. *Projection of $W^u(\boldsymbol{z^0})$ onto the (ψ_1, ψ_2) plane ($\Sigma = 15.5$; coloring indicates σ; a total of $77,500$ mesh-points) computed* with *(left) and* without *(right) "projection onto the energy level" procedure. Black lines indicate the "cell" and homoclinic/heteroclinic trajectories. The light green "ink spots" indicate regions unattainable due to the energy conservation (as in Figure 11). The right picture clearly shows the loss of energy in the process of computation.*

unstable manifolds $W^u(\boldsymbol{z^1})$ and $W^u(\boldsymbol{z^{-1}})$; see Figure 12B. This "folding onto itself" is a numerical artifact rather than a feature of $W^u(\boldsymbol{z^0})$.

To handle this problem, we implement an additional step of projection onto the energy level: Immediately before a *Considered* mesh-point $\boldsymbol{y}$ is *Accepted*, we solve an initial value problem

$$\boldsymbol{y}_p'(t) = \text{sign}\,(2 - E(\boldsymbol{y}))\,\nabla E(\boldsymbol{y}_p(t)), \qquad \boldsymbol{y}_p(0) = \boldsymbol{y},$$

until the first intersection $\boldsymbol{y}_p(\tau)$ with the level set $E = 2$. If that point is within $(\Delta/10)$ from $\boldsymbol{y}$, we set $\boldsymbol{y} = \boldsymbol{y}_p(\tau)$ and continue as described in section 6; otherwise, the algorithm terminates since the local energy loss after solving the PDE is considered too large.

This results in a much better approximation of $W^u(\boldsymbol{z^0})$; see Figure 12A. Unfortunately, this projection procedure becomes unstable near all $\boldsymbol{z^m}$'s since both $W^u(\boldsymbol{z^m})$ and $W^s(\boldsymbol{z^m})$ lie in the same energy level $E = 2$, which becomes singular at each $\boldsymbol{z^m}$. Thus, our implementation artificially stops the manifold from growing too close to those points; i.e., tentative triangles are not added to segments of *AcceptedFront* which are within $R_{restrict} = 0.3$ from $\boldsymbol{z^m}$.

Figure 13 shows two homoclinic orbits of $\boldsymbol{z^0}$ and two heteroclinic orbits connecting $\boldsymbol{z^0}$ to $\boldsymbol{z^1}$ and $\boldsymbol{z^{-1}}$. These trajectories appear to lie on the computed manifold-approximation, indirectly confirming convergence to $\mathrm{W}^u_\Sigma(\boldsymbol{z^0})$. However, a direct verification of convergence for this example is hard due to the lack of analytic formulae for $W^u(\boldsymbol{z^0})$.

9. Example: An egg carton surface. Finally, in order to test the rate of convergence numerically, we consider a simple example for which the invariant manifold is a priori known.

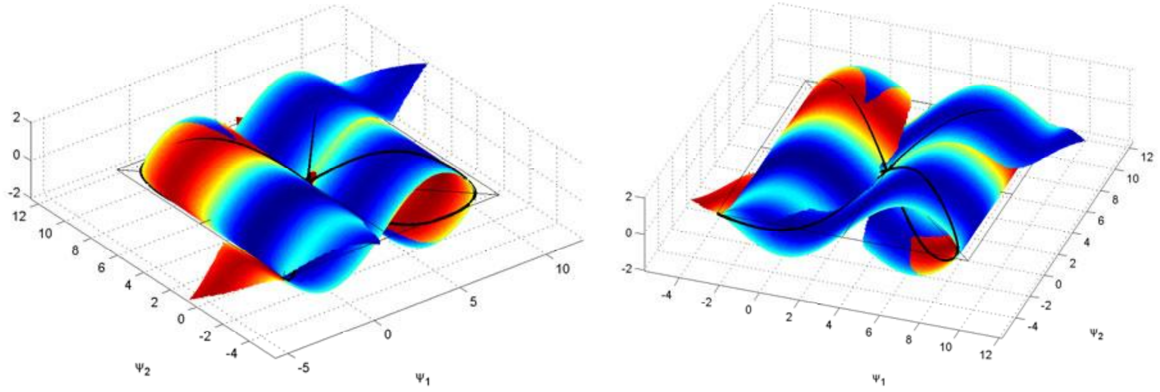


Figure 13. *Two different (rotated) views of the projection of $W^u(\boldsymbol{z^0})$ into the $(\psi_1, \psi_2, \psi_1')$ space. Conservation of energy is enforced by the projection procedure. Coloring is used to indicate the fourth coordinate (ψ_2') ranging from -2.07 (blue) to 2.07 (red). The seeming self-intersection is a side effect of the three-dimensional projection side effect. The thin black lines indicate the "cell." The thick black lines indicate sample homoclinic and heteroclinic trajectories.*

Given a smooth function $g(x, y)$, we consider a system

$$\begin{aligned} x' &= \eta_1 x; \\ y' &= \eta_2 y; \\ z' &= -\mu z + \mu g(x, y) + \eta_1 x g_x(x, y) + \eta_2 y g_y(x, y). \end{aligned} \tag{17}$$

If η_1, η_2, and μ are positive, then the point $(0, 0, g(0, 0))$ is a saddle and the graph of $g(x, y)$ is its unstable manifold. For testing purposes, we have chosen an "egg carton" function $g(x, y) = 0.27 \sin(2\pi x) \sin(2\pi y)$; see Figure 14. Of course, the choice of (η_1, η_2, μ) also influences the computational error; e.g., a bigger μ will obviously make this an easier problem since μ is the rate at which all trajectories are pushed toward the manifold.

For every mesh-point (x, y, z), the approximation error $\mathcal{E}$ is the distance to the manifold surface. An upper bound is readily available as $\mathcal{E}(x, y, z) \leq |z - g(x, y)|$ and can be used to compute the bound on L_2 and L_∞ errors for the entire mesh. In these tests we used $\varSigma = 1.5$ and two different values for μ (1 and 1/4). All computations were repeated for the "isotropic" case $\eta_1 = \eta_2 = 1$ (see Figure 15) and for the "anisotropic" case $\eta_1 = \eta_2/5 = 1$ (see Figure 16). As expected, we observe a quadratic growth of the number of mesh-points N and a linear decay of the approximation error in all of the examples.

10. Conclusions. We have introduced a fast algorithm for approximating invariant manifolds of saddle points of the vector fields in R^n. The chief advantage of this method is its efficiency: all the examples presented in sections 7 and 8 take under 90 seconds to compute on a Pentium III 850 MHz processor with 256Mb RAM. Our approach is new and many related issues remain open. Possible directions for future work include higher-order methods, error bounds and estimates (possibly using an interval arithmetic implementation), adaptive and parallel methods, exploration of robustness under parameter variation, and proofs of convergence. The previously available methods described in section 2 are more developed

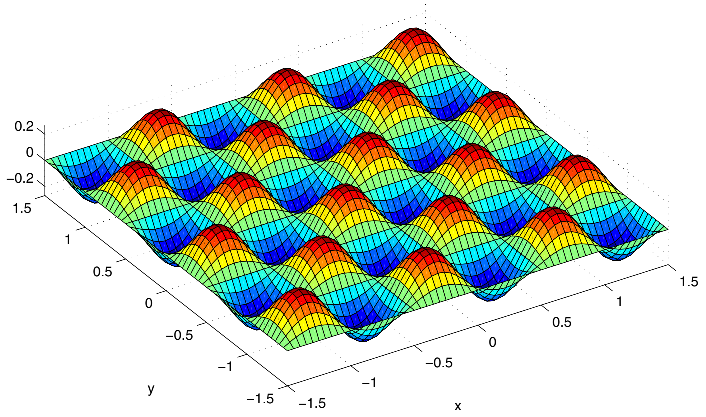
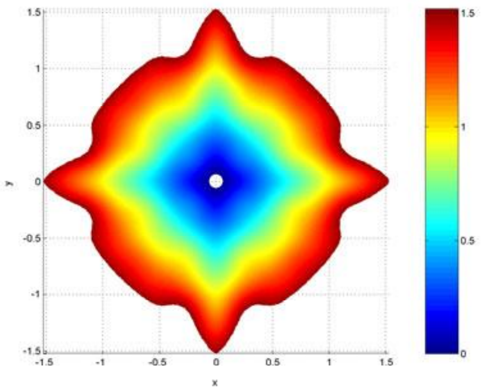


Figure 14. *An "egg carton" function* $g(x, y) = 0.27\sin(2\pi x)\sin(2\pi y)$.



Parameters		$\mu = 1$		
Δ	R_{init}	N	L_2 Error	L_∞ Error
$\Delta_0 = 0.05$	$r_0 = 0.2$	3940	0.034316	0.103143
$\Delta_0 \times 2^{-1}$	$r_0 \times 2^{-1}$	14039	0.016671	0.056432
$\Delta_0 \times 2^{-2}$	$r_0 \times 2^{-2}$	53271	0.008582	0.028589
$\Delta_0 \times 2^{-3}$	$r_0 \times 2^{-3}$	208626	0.004168	0.015052
$\Delta_0 \times 2^{-4}$	$r_0 \times 2^{-4}$	829158	0.002086	0.007570

Parameters		$\mu = 0.25$		
Δ	R_{init}	N	L_2 Error	L_∞ Error
$\Delta_0 = 0.05$	$r_0 = 0.2$	3894	0.046520	0.152055
$\Delta_0 \times 2^{-1}$	$r_0 \times 2^{-1}$	13936	0.023853	0.081145
$\Delta_0 \times 2^{-2}$	$r_0 \times 2^{-2}$	53159	0.012685	0.043746
$\Delta_0 \times 2^{-3}$	$r_0 \times 2^{-3}$	208502	0.005986	0.021406
$\Delta_0 \times 2^{-4}$	$r_0 \times 2^{-4}$	828884	0.002969	0.011011

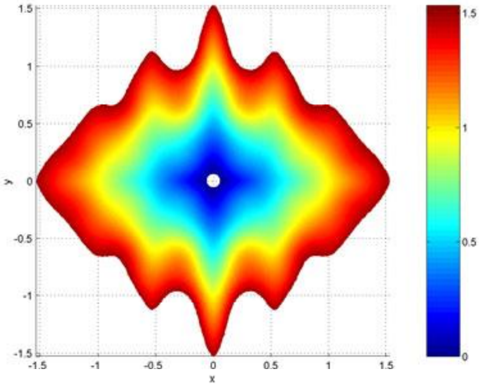
Figure 15. *Isotropic case ($\eta_1 = 1$, $\eta_2 = 1$). The (x-y)-projection of the mesh (color indicates trajectory-arc-length σ) and the table of error bounds.*

(with many extensions available), but, to the best of our knowledge, are substantially more time-consuming on the problems with multiple time-scales.

The perspective of building the manifold as a collection of simplexes, each of them satisfying a locally posed PDE, is quite general. For example, customized stopping criteria can be used to treat manifolds converging to attracting limit sets. We are also planning to investigate the applicability of our approach to approximating invariant manifolds of saddle-type cycles (see [19] and [26] for the existing methods).

Our current implementation relies on $k = 2$ for the local mesh-generation procedure only. We expect that a combination of our approach with robust techniques for higher-dimensional mesh extension will yield fast methods for the general case (see Remark 6.5).

Finally, we note that some of the ideas illustrated above may be useful in the context



Parameters		$\mu = 1$		
Δ	R_{init}	N	L_2 Error	L_∞ Error
$\Delta_0 = 0.05$	$r_0 = 0.2$	3629	0.038158	0.090629
$\Delta_0 \times 2^{-1}$	$r_0 \times 2^{-1}$	13429	0.018711	0.049707
$\Delta_0 \times 2^{-2}$	$r_0 \times 2^{-2}$	51466	0.009302	0.024855
$\Delta_0 \times 2^{-3}$	$r_0 \times 2^{-3}$	204446	0.004601	0.011553
$\Delta_0 \times 2^{-4}$	$r_0 \times 2^{-4}$	815410	0.002374	0.006496

Parameters		$\mu = 0.25$		
Δ	R_{init}	N	L_2 Error	L_∞ Error
$\Delta_0 = 0.05$	$r_0 = 0.2$	3621	0.040681	0.127953
$\Delta_0 \times 2^{-1}$	$r_0 \times 2^{-1}$	13425	0.020136	0.059774
$\Delta_0 \times 2^{-2}$	$r_0 \times 2^{-2}$	52066	0.010428	0.026192
$\Delta_0 \times 2^{-3}$	$r_0 \times 2^{-3}$	205321	0.005296	0.013102
$\Delta_0 \times 2^{-4}$	$r_0 \times 2^{-4}$	815905	0.002631	0.007336

Figure 16. *Anisotropic case ($\eta_1 = 1$, $\eta_2 = 5$). The (x-y)-projection of the mesh (color indicates trajectory-arc-length σ) and the table of error bounds.*

of prior methods, which grow the manifold as a collection of $(k-1)$-dimensional topological spheres. In particular, we believe that the method defined in [20, 21] can be substantially accelerated by using parts of M_{i+1} as they become available (as opposed to using M_i only and producing the entire M_{i+1} at once). Further speedup can be attained by ordering the computation of markers (first compute those whose trajectories are "the least tangential" to M_i) and using a discretized system-solver instead of the time-consuming shooting methods.

Acknowledgments. The authors would like to thank S. Vavasis, K. Lin, O. Junge, H. Osinga, B. Krauskopf, and R. Sacker.

REFERENCES

[1] R. Altendorfer, R. 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 and the Support Technologies for Mobile Machines (CLAWAR 2002), P. Bidaud and F. Ben Amar, eds., Professional Engineering Publishing, London, UK, 2002. Available online at http://ai.eecs.umich.edu/CNM/Publications/Published_Articles/Clawar2002.pdf

[2] H. D. Chiang, F. F. Wu, and P. P. Varaiya, *A BCU method for direct analysis of power system transient stability*, IEEE Trans. Power Systems, 9, (1994), pp. 1194–1208.

[3] D. G. Aronson, E. J. Doedel, J. Guckenheimer, and B. Sandstede, *On the Dynamics of Torsion-Coupled Pendula*, in preparation.

[4] M. L. Brodzik, *The computation of simplicial approximations of implicitly defined p-dimensional manifolds*, Comput. Math. Appl., 36, (1998), pp. 93–113.

[5] Paul L. Chew, *Guaranteed-Quality Triangular Meshes*, Tech. report 89-983, Department of Computer Science, Cornell University, Ithaca, NY, 1989.

[6] 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, 1996, pp. 449–459.

[7] M. Dellnitz and A. Hohmann, *A subdivision algorithm for the computation of unstable manifolds and global attractors*, Numer. Math., 75 (1997), pp. 293–317.

[8] L. Dieci and J. Lorenz, *Computation of invariant Tori by the method of characteristics*, SIAM J. Numer. Anal., 32 (1995), pp. 1436–1474.

[9] L. Dieci, J. Lorenz, and R. D. Russell, *Numerical calculation of invariant tori*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 607–647.

[10] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov,, B. Sandstede, and X. J. Wang, *AUTO97 : Continuation and Bifurcation Software for Ordinary Differential Equations*, 1997. Available online at ftp://ftp.cs.concordia.ca/pub/doedel/auto/auto.ps.gz

[11] E. J. Doedel, *Private communication*, IMA, October 1997.

[12] K. D. Edoh, R. D. Russell, and W. Sun, *Orthogonal Collocation for Hyperbolic PDEs and Computation of Invariant Tori*, Mathematics Research Report MRR 060-95, Australian National University, Canberra, Australia, 1995.

[13] N. Fenichel, *Persistence and smoothness of invariant manifolds for flows*, Indiana Univ. Math. J., 21, (1971), pp. 193–226.

[14] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*, Springer-Verlag, Berlin, 1983.

[15] J. Guckenheimer and P. Worfolk, *Dynamical systems: Some computational problems*, in Bifurcations and Periodic Orbits of Vector Field, D. Schlomiuk, ed., NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 408, Kluwer Academic Publishers, 1993.

[16] M. E. Henderson, *Computing invariant manifolds by integrating fat trajectories*, SIAM J. Applied Dynamical Systems, submitted.

[17] M. E. Henderson, *Multiple parameter continuation: Computing implicitly defined k-manifolds*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 451–476.

[18] 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.

[19] 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, 11 (2001), pp. 1–18.

[20] B. Krauskopf and H. Osinga, *Two-dimensional global manifolds of vector fields*, Chaos, 9 (1999), pp. 768–774.

[21] B. Krauskopf and H. Osinga, *Global Manifolds of Vector Fields: The General Case*, Applied Nonlinear Mathematics Research Report 99.2, University of Bristol, Bristol, UK, 1999.

[22] B. Krauskopf and H. Osinga, *Visualizing the structure of chaos in the Lorenz system*, Comput. Graphics, 26 (2002), pp. 815–823.

[23] E. N. Lorenz, *Deterministic nonperiodic flows*, J. Atmospheric Sci., 20 (1963), pp. 130–141.

[24] H. Mingyou, T. Küpper, and N. Masbaum, *Computation of invariant tori by the Fourier methods*, SIAM J. Sci. Comput., 18 (1997), pp. 918–942.

[25] J. Moser, *On invariant manifolds of vector fields and symmetric partial differential equations*, Differential Anal., Bombay Colloq., (1964), pp. 227–236.

[26] H. Osinga, *Nonorientable manifolds of three-dimensional vector fields*, Internat. J. Bifur. Chaos, 13 (2003), pp. 553–570.

[27] J. Peraire, J. Peiro, and K. Morgan, *Advancing front grid generation*, in Handbook of Grid Generation, J. F. Thompson, B. K. Soni, and N. P. Weatherill, eds., CRC Press, Boca Raton, FL, 1999, Chapter 17.

[28] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C,* Cambridge University Press, Cambridge, UK, 1988.

[29] S. Rebay, *Efficient unstructured mesh generation by means of Delaunay triangulation and Bowyer-Watson algorithm*, J. Comput. Phys., 106 (1993), pp. 125–138.

[30] R. J. Sacker, *A new approach to the perturbation theory of invariant surfaces*, Comm. Pure Appl. Math., 18 (1965), pp. 717–732.

[31] S. Fomel and J. A. Sethian, *Fast-phase space computation of multiple arrivals*, Proc. Nat. Acad. Sci., 99 (2002), pp. 7329–7334.

[32] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton-Jacobi equations*, Proc. Nat. Acad. Sci., 98 (2001), pp. 11069–11074.

[33] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton–Jacobi equations: Theory and applications*, SIAM J. Numer. Anal., 41 (2003) pp. 325–363.

[34] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for hybrid control*, in Proceedings of the 5th International Workshop (HSCC 2002, Stanford, CA, March 25-27, 2002), Lecture Notes in Comput. Sci. 2289, Springer-Verlag, New York, NY, pp. 393–406.

[35] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1994.

[36] A. Vladimirsky, *Space-Marching Methods for the First-Order Non-Linear PDEs*, in preparation.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 3, pp. 261–351

Analysis of Rotation–Vibration Relative Equilibria on the Example of a Tetrahedral Four Atom Molecule[*]

K. Efstathiou[†], D. A. Sadovskii[†], and B. I. Zhilinskii[†]

Abstract. We study relative equilibria (RE) of a nonrigid molecule, which vibrates about a well-defined equilibrium configuration and rotates as a whole. Our analysis unifies the theory of rotational and vibrational RE. We rely on the detailed study of the symmetry group action on the initial and reduced phase space of our system and consider the consequences of this action for the dynamics of the system. We develop our approach on the concrete example of a four-atomic molecule A_4 with tetrahedral equilibrium configuration, a dynamical system with six vibrational degrees of freedom. Further applications and illustrations of our results can be found in [van Hecke et al., *Eur. Phys. J. D At. Mol. Opt. Phys.*, 17 (2001), pp. 13–35].

Key words. small vibrations, vibration-rotation of molecules, spherical top, relative equilibria, 1:1:1 resonant oscillator, normalization, reduction, bifurcations, orbit space, finite group action, reversing symmetry, Molien generating function, integrity basis

AMS subject classifications. 37J15, 37J35, 37J40, 81V55, 58D19

DOI. 10.1137/030600015

1. Introduction.

This paper unifies modern methods of classical theory of symmetric Hamiltonian dynamical systems and quantum theory of molecules (and other isolated finite-particle systems). Considerable progress was achieved in both directions in the last decades and deep relations between these seemingly distant theories became evident. Significant effort by mathematicians and molecular physicists to converge the two fields resulted in the qualitative theory of highly excited quantum molecular systems based on recent mathematical developments. We join the two approaches and demonstrate what kind of concrete results can be immediately obtained in molecular systems [1, 2, 3, 4] by applying powerful methods of symmetric Hamiltonian systems [5, 6, 7, 8, 9, 10, 11]. We choose a concrete problem of rotation–vibration of a four-atomic molecule with tetrahedral equilibrium configuration [12, 13] in order to explain the details of our approach.

1.1. Vibrational relative equilibria or nonlinear normal modes.

Montaldi, Roberts, and Stewart [14, 15, 16] gave a general description of periodic solutions near equilibria of symmetric Hamiltonian systems: the so-called nonlinear normal modes or relative equilibria (RE). They related the number of RE to the symmetry group of the system and showed, on several examples of bound systems of vibrating particles, that this number can be significantly larger than the number of vibrational degrees of freedom. This mathematical result was not fully

[*] Received by the editors January 30, 2003; accepted for publication (in revised form) by D. Holm October 30, 2003; published electronically July 6, 2004. This work was supported by the EU project Mechanics and Symmetry in Europe (MASIE), contract HPRN-CT-2000-00113.

http://www.siam.org/journals/siads/3-3/60001.html

[†] Université du Littoral, UMR 8101 du CNRS, 59 140 Dunkerque, France (konstantinos@efstathiou.gr, sadovski@univ-littoral.fr, zhilin@univ-littoral.fr).

appreciated by molecular physicists until it was reproduced by an alternate technique [17, 18, 19] based on the analysis of the reduced system in the so-called polyad approximation, a generalization of the approximation used for two-oscillator systems in [20, 21, 22, 23, 24, 25, 26, 27] and others. It was shown that fixed points of the symmetry group action on the reduced phase space correspond to vibrational RE. Later work [28, 29, 30][1] uncovered more fully the correspondence of both approaches and bridged the differences in their tools and terminology.

1.2. Rotational RE or stationary axes of rotation.

Similar analysis of stationary points of the reduced rotational system [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43] was initiated in molecular physics even before the analysis of vibrational RE. In terms of symmetric Hamiltonian systems [45, 46, 44], this analysis is equivalent to studying rotational RE [28, 29]. This was demonstrated in the recent study [12] of the rotational structure of the tetrahedral molecule P_4 [47],[2] where the energy of rotational RE is derived from the parameters of the internuclear potential. Our present analysis extends the method in [12] to rotation–vibration systems.

1.3. Applications in molecular physics and spectroscopy.

Classical analysis of different kinds of RE is used for the description of molecular energy level spectra on the basis of the classical quantum correspondence principle, which links the topological description of the classical dynamical system to such qualitative aspects of quantum spectra as existence of bands, polyads, clusters, and their persistence under small modifications of parameters. Some of these qualitative characteristics are discussed in the present paper. Review articles [48, 49, 50, 51, 52, 53][3] give more examples of molecular applications and initiation to formal theory.

Much of the work in molecular spectroscopy is done using so-called effective model Hamiltonians $H_{\rm eff}$, which describe explicitly only a fraction of degrees of freedom of the system and treat other degrees effectively. In other words, $H_{\rm eff}$ describe reduced systems, where reduction is based on a model assumption of approximate separability and/or approximate dynamical symmetries. Equilibria (stationary points) of $H_{\rm eff}$ are RE of the initial system.

In practice, reduction often remains only an abstract theoretical possibility because parameters in the full initial molecular Hamiltonian are unknown. So, parameters of $H_{\rm eff}$ are simply fitted to experimental data. Classical analogues of such phenomenological model Hamiltonians can be constructed if excitation is sufficiently high to validate the classical limit. When, as it is often the case, only some degrees of freedom described by $H_{\rm eff}$ (e.g., rotation) can be meaningfully treated as classical, the rest (e.g., vibration) is kept quantum. The energies of such hybrid “semiquantum” systems are eigenvalues that depend on the dynamical variables of the classical subsystem. The most well-known example of semiquantum energy is the rotational energy surfaces of vibration-rotation systems [36, 35, 41].

[1] The approach of Montaldi and Roberts is less oriented to the reduced problem and thus can be potentially extended to molecules in which separation of vibration and rotation and the introduction of the molecule-fixed frame is problematic. For the relatively rigid molecules, we consider their approach as being equivalent to ours.

[2] Among the few different molecules of type A_4, the phosphorus P_4 is studied experimentally; see [47].

[3] Our approach follows closely the ideas in [48, 49, 50], which review group actions and their applications in physics.

The semiquantum approach turned out to be very fruitful, and numerous vibration-rotation systems at low vibrational excitation were analyzed in great detail [37, 38, 18, 54, 55, 56, 42, 57, 43, 58]. In particular, typical (universal) modifications of the cluster structure of the energy level spectrum, or quantum bifurcations, were described in terms of modifications of the set of stationary points of the energy surfaces. Direct, explicit relation of these stationary points to classical RE was established recently in [58, 13]. Other important qualitative quantum phenomena include rearrangements (crossings) of energy level bands [59, 43, 60] and quantum monodromy [61, 62], which are interpreted as crossings of semiquantum energies and are also related to classical RE [60, 63, 64, 65].

1.4. Main idea. We combine recent theories of rotational RE and vibrational nonlinear normal modes in order to study the rotation–vibration problem. Using symmetry and topology, we find particular solutions (critical orbits) common to a whole class of model systems with given symmetry and with different potentials. Subsequently, we define a concrete potential, normalize the classical system, and construct explicitly the effective Hamiltonian H_{eff}. Using this Hamiltonian, we obtain quantitative predictions for concrete molecular models, which illustrate general qualitative results. We explain our approach in the example of rotation–vibration of the four-atomic homonuclear molecule A_4 with tetrahedral equilibrium configuration [47].

2. Basic aspects of the analysis. We review certain general definitions, which are used later in the paper, and give the plan of the analysis.

2.1. Symmetry group T_d and its extensions. Along with translation and rotational symmetry, which are present for any isolated finite-particle system in the absence of external fields, each molecule possesses its own internal symmetry related to the existence of identical particles. The symmetry group of our system originates from the spatial symmetry group T_d of the tetrahedral equilibrium configuration of A_4 and momentum reversal $\mathcal{T}$, which in the original system sends (q, p) to $(q, -p)$, and is discussed in more detail in section 3.1. Our initial Hamiltonian is invariant with respect to these symmetries. As an abstract group, T_d is the permutation group of four identical objects. We use the Schönflis point group notation [67, 66], which is standard in molecular physics. Irreducible representations of T_d are most frequently labeled in molecular physics as A_1, A_2 (one-dimensional), E (two-dimensional), and F_1, F_2 (three-dimensional).

2.2. Vibrational degrees of freedom of an A_4 molecule. The A_4 molecule has six vibrational degrees of freedom, which constitute the nondegenerate "breathing" mode A_1, and the doubly and triply degenerate modes E and F_2. The spectroscopic notation of these modes is $\nu_1^{A_1}$, ν_2^{E}, $\nu_3^{F_2}$. We use a simplified notation for the coordinates and conjugate momenta of the modes given in Table 1 and we also use classical complex oscillator variables

$$z = q + ip, \quad \bar{z} = q - ip. \tag{2.1}$$

The zero order vibrational Hamiltonian H_0 of A_4 represents a 1-oscillator, a 1:1 oscillator, and a 1:1:1 oscillator with frequencies ω_{A_1}, ω_E, and ω_{F_2}, respectively.

2.3. Rotation–vibration Hamiltonian. Assuming that the static equilibrium configuration of A_4 about which the atoms are vibrating is well defined and the amplitudes of vibrations

Table 1

Notation for vibrational and rotational dynamical variables of the A_4 *molecule. Expression of angular momenta* j_α *in terms of dynamical variables of a two-dimensional oscillator (Schwinger representation) is used in this table and throughout the paper.*

Subsystem	Traditional notation	This paper
F_2 mode	$q_\alpha^{F_2}$, $p_\alpha^{F_2}$, $\alpha = x, y, z$	q_i, p_i, $i = 1, 2, 3$
E mode	q_α^E, p_α^E, $\alpha = a, b$ or $1, 2$	q_i, p_i, $i = 4, 5$
A_1 mode	q^{A_1}, p^{A_1}	q_a, p_a
rotation	momenta j_α, $\alpha = x, y, z$	q_i, p_i, $i = 6, 7$

are small, we can separate molecular rotation and define the frame rotating with the molecule. The molecule is isolated, external fields are absent, and translation of the center of mass is therefore excluded.

To derive the rotation–vibration Hamiltonian H, we can follow the procedure described in Chapter 11 of [1] and, more rigorously, in Chapter 7.10 of [68]. The molecule-fixed frame is related to the equilibrium configuration by the Eckart conditions. The kinetic energy T is a complicated function

$$2T = \sum_i m_i \left[(\Omega \wedge (\mathbf{R}_i^0 + \mathbf{r}_i))^2 + \dot{\mathbf{r}}_i^2 + 2\Omega(\mathbf{r}_i \wedge \dot{\mathbf{r}}_i) \right] \tag{2.2}$$

of small vibrational displacement velocities and angular velocities defined with respect to this frame. The intramolecular potential U can be simply written in terms of vibrational coordinates. The Hamiltonian form requires rotational angular momenta j, defined in the molecule-fixed frame, and vibrational coordinates q and momenta p.

To put the initial Hamiltonian in the form suitable for normalization, we Taylor expand $H = T(q, p, j) + U(q)$ in q and rescale (p, q) to bring the harmonic part to the standard form. We then express the components of $\mathbf{j}$ in terms of coordinates and momenta of the auxiliary two-dimensional harmonic oscillator in order to treat vibrational and rotational variables in the same way and use complex variables z in (2.1). The resulting formal power series expression

$$H = \omega(H_0 + \epsilon H_1 + \epsilon^2 H_2 + \epsilon^3 H_3 + \cdots) \tag{2.3}$$

is the starting point of the normal form transformation.

We use the concrete example of the phosphorus molecule P_4 with the tetrahedral equilibrium configuration and harmonic atom–atom bond potential [13] to illustrate our results. The only two molecular parameters in this example are the energy scale ω and the dimensional smallness parameter $\epsilon = (kmr)^{-1}$ in the series expansion. Here r, m, and k stand for the interatomic distance, the mass of the atoms, and the force constant of the potential, respectively. The values of ϵ and ω can be used as phenomenological parameters to reproduce experimental data qualitatively: $\epsilon \approx 2 \times 10^{-2}$ and $\omega \approx 329$ cm^{-1} for P_4 [13].

2.4. Reduced system. The approximate dynamical symmetry of the system with Hamiltonian (2.3) is defined by the zero order term H_0. We suppose that the frequencies ν_{A_1}, ν_E, and ν_{F_2} are incommensurate; i.e., we assume the absence of any resonances between different vibrational modes. In such a case, we can introduce reduced phase spaces for each of the subsystems simultaneously. The total reduced phase space is the product of these spaces. The normal form H_{eff} is an effective rotation–vibration Hamiltonian describing polyads of nonresonant modes A_1, E, and F_2. For simplicity, we neglect the A_1 mode (i.e., we set $q^{A_1} = p^{A_1} = 0$) and focus on modes E and F_2.

2.4.1. Rotational subsystem; rotational space $\mathbb{S}^2$. Conservation of the total angular momentum is the consequence of the isotropy of physical space (in the absence of external fields). The rotational dynamical variables j_α ($\alpha = 1, 2, 3$) are subjected to the constraint $j_1^2 + j_2^2 + j_3^2 =$ const and the rotational phase space is a two-dimensional sphere $\mathbb{S}^2$, which can be constructed in the space $\mathbb{R}^3$ with coordinates j_α. For the auxiliary two-dimensional oscillator, used to represent the momenta j_α (Table 1), the restriction $\mathbf{j}^2 =$ const is equivalent to fixing the sum of two actions.

2.4.2. E-mode subsystem, vibrational space $\mathbb{C}P^1 \sim \mathbb{S}^2$. Exploiting the well-known equivalence of the two-dimensional 1:1 harmonic oscillator and an angular momentum system, we introduce vibrational angular momenta v_1, v_2, v_3 [69, 20, 21]. The internal structure of vibrational polyads formed by the doubly degenerate vibrational mode E can be described in terms of these dynamical variables. The E-mode polyad sphere $\mathbb{S}^2$ is defined by the equation

$$v_1^2 + v_2^2 + v_3^2 = n_e^2 = \text{const}$$

in the ambient space $\mathbb{R}^3$ with coordinates (v_1, v_2, v_3). Any point on this sphere is uniquely represented by the values of (v_1, v_2, v_3) if we keep in mind that $v_1^2 + v_2^2 + v_3^2$ is a constant. The diffeomorphic space $\mathbb{C}P^1$ can be defined in $C_2 - \{0\}$ using the equivalence class of points z_4:z_5. Two complex numbers (z_4, z_5) can be used as coordinates on $\mathbb{C}P^1_{n_e}$ if their modules are restricted as

$$|z_4|^2 + |z_5|^2 = 2n_e$$

and all pairs (z_4, z_5), which differ in a common phase factor $e^{i\phi}$, correspond to the same point of $\mathbb{C}P^1$. For example, coordinates $(v_1, v_2, v_3) = (0, 1, 0)$ and $(z_4, z_5) = (1, -i) = (e^{i\phi}, e^{i\phi - \pi/2})$ define the same point.

2.4.3. F_2-mode subsystem; vibrational space $\mathbb{C}P^2$. Generalization of the above construction for the F_2-mode 1:1:1 oscillator [17, 18, 70] leads to the reduced phase space $\mathbb{C}P^2_{n_f}$. The approximate integral of motion equals

$$\tfrac{1}{2}(z_1\bar{z}_1 + z_2\bar{z}_2 + z_3\bar{z}_3) = n_f \approx \text{const},$$

and (z_1, z_2, z_3) can be used as coordinates on $\mathbb{C}P^2$. In fact, we can define a point on the $\mathbb{C}P^k$ space as an equivalence class of points on C_{k+1} given by their homogeneous coordinates $z_1 : z_2 : \cdots : z_{k+1}$ or, equivalently, as a class of points on $(z_1, z_2, \ldots, z_{k+1}) \in C_{k+1}$ defined up to a common phase $(z_1, z_2, \ldots, z_{k+1}) \sim e^{i\phi}(z_1, z_2, \ldots, z_{k+1})$ and such that $|z_1|^2 + |z_2|^2 + \cdots + |z_{k+1}|^2$ is a constant.

2.4.4. Full reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. Before reduction, our initial molecular system has five vibrational degrees of freedom (if the nondegenerate A_1 mode is neglected) and two auxiliary oscillatory degrees of freedom introduced to describe the rotational subsystem. Three independent reductions fix the strict integral of motion j (the amplitude of the total angular momentum) and polyad integrals n_e and n_f of the doubly degenerate mode E and the triply degenerate mode F_2. The reduced system is left with only four degrees of freedom. Reduction makes the topology of the reduced phase space more complicated. The total reduced space is a direct product of the rotational phase sphere, $\mathbb{S}^2_j$, E-mode vibrational polyad sphere $\mathbb{C}P^1_{n_e} \sim \mathbb{S}^2$, and F_2-mode vibrational polyad phase space $\mathbb{C}P^2_{n_f}$. Omitting extra indexes and shortening the notation, we represent the topology of the reduced phase space simply as $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, where $\mathbb{S}^2$ and $\mathbb{C}P^1$ stand for the rotational and vibrational E-mode phase spaces, respectively.

2.4.5. Normal form. Once the Hamiltonian function H is in the oscillator form (2.3), we can normalize it using the standard Lie transform method [71, 72, 73, 74]. All odd orders [odd degrees in $(z, \bar{z})$] vanish in the normal form

$$\mathcal{H}_{\mathrm{nf}} = \omega(H_0 + \epsilon^2\mathcal{H}_2 + \epsilon^4\mathcal{H}_4 + \epsilon^6\mathcal{H}_6 + \cdots), \tag{2.4}$$

which is a power series in ϵ^2. To obtain the reduced Hamiltonian H_{eff}, the terms $\mathcal{H}_{2k}$ in (2.4) should be expressed as functions of basic invariant polynomials (of all generators of the algebra of invariant polynomials) on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. Due to algebraic dependencies between generators (or "sygyzies") a special polynomial basis should be constructed. A general solution to this problem is provided by a Gröbner basis. Two more specialized polynomial bases—an integrity basis used in invariant theory, and a tensorial basis used by spectroscopists to represent effective Hamiltonians—can be used. We further discuss these bases in section 6.4.

2.5. Scheme of the analysis. Our analysis of a finite-particle quantum system includes several steps: (i) construction of the initial complete classical Hamiltonian H and of the corresponding quantum operator; (ii) reduction of H, taking into account strict and approximate integrals of motion, i.e., the "model"; (iii) analysis of classical RE, relative periodic orbits, and invariant submanifolds;

(iv) interpretation in terms of quantum energy spectrum. Each step has a general part and a concrete part. Many important general results follow from the topology of the reduced phase space and the symmetry group action on it, i.e., from the model.

In the first half of the paper, which includes sections 4 and 5, we find as much information about our system as possible before any concrete interaction potential bounding the particles is introduced explicitly and even before any dynamics is studied. After establishing the topology of the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ and the invariance symmetry group $T_d \times \mathcal{T}$ of our system, we study the action of the group $T_d \times \mathcal{T}$ on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. To this end, we first consider the action on the individual factor spaces $\mathbb{C}P^2$, $\mathbb{C}P^1$, and $\mathbb{S}^2$ and then extend it to the full reduced space. Time reversal $\mathcal{T}$ and other reversing symmetries, which include $\mathcal{T}$, are antisymplectic and should be treated differently from purely spatial symplectic symmetries.

We assume that the reduced Hamiltonian H_{eff} is a generic $T_d \times \mathcal{T}$ invariant Morse function on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. The RE of our system are stationary points of H_{eff}, which exist anywhere

Table 2

Classes of conjugated and invariant subgroups of the $T_d \times \mathcal{T}$ group. Part I.

	Class[4]	Structure[5]	Description and comments
	Subgroups of order 1		
1	C_1	$\{1\}$	Trivial subgroup.
	Subgroups of order 2		
1	$\mathcal{T}$	$\{1, \mathcal{T}\}$	Momentum (or time) reversal, also denoted as Z_2.
6	C_s	$\{1, \sigma^\alpha\}$	Reflection in a plane, T_d has six conjugated operations σ^α.
6	$\mathcal{T}_s$	$\{1, \mathcal{T}_s\}$	Simultaneous reflection and time reversal,[8] also denoted as $(\sigma\mathcal{T})$ or (σZ_2).
3	C_2	$\{1, C_2\}$	Rotation by π around one of axes S_4^a, $a = (x, y, z)$.
3	$\mathcal{T}_2$	$\{1, \mathcal{T}_2\}$	Rotation by π and time reversal,[6] also denoted as $(C_2\mathcal{T})$ or (C_2Z_2).
	Subgroups of order 3		
3	C_3	$\{1, C_3, C_3^2\}$	Cyclic rotational subgroups corresponding to four different C_3 axes of the T_d group.
	Subgroups of order 4		
1	D_2	$\{1, C_2^x, C_2^y, C_2^z\}$	An invariant subgroup of the T_d group.
3	S_4	$\{1, S_4, C_2, S_4^3\}$	Cyclic groups generated by the S_4^a operations $a = (x, y, z)$.
3	$\mathcal{T}_4$	$\{1, \mathcal{T}_{+4}, C_2, \mathcal{T}_{-4}\}$	Cyclic groups generated by the $S_4^a \circ \mathcal{T} = \mathcal{T}_{+4}^a$ operations.[9]
3	$C_2 \times \mathcal{T}_2$	$\{1, C_2^a, \mathcal{T}_2^b, \mathcal{T}_2^c\}$	(a, b, c) is one of the three cyclic permutations of (x, y, z).[6]
6	$C_s \times \mathcal{T}_2$	$\{1, \sigma^{a_1}, \mathcal{T}_2^a, \mathcal{T}_s^{a_2}\}$	These correspond to six different choices of the σ^{a_1} symmetry plane.[6,7,8]
3	C_{2v}	$\{1, C_2^a, \sigma^{a_1}, \sigma^{a_2}\}$	Subgroup of T_d. Axis a is one of (x, y, z).[7]
3	$C_2 \times \mathcal{T}_s$	$\{1, C_2^a, \mathcal{T}_s^{a_1}, \mathcal{T}_s^{a_2}\}$	Obtained from C_{2v} by combining two reflections and time reversal[7,8]
3	$C_2 \times \mathcal{T}$	$\{1, C_2, \mathcal{T}, \mathcal{T}_2\}$	direct product of C_2^a and time reversal. Axis a is one of (x, y, z)[6]
6	$C_s \times \mathcal{T}$	$\{1, \sigma, \mathcal{T}, \mathcal{T}_s\}$	corresponding to one of the six conjugated symmetry planes σ of the T_d group.[8]
	Subgroups of order 6		
4	$C_3 \times \mathcal{T}$	$\{1, C_3, C_3^2,$ $\mathcal{T}, C_3\mathcal{T}, C_3^2\mathcal{T}\}$	Direct product of C_3 and time reversal $\mathcal{T}$ corresponding to four different axes C_3.
4	C_{3v}	$\{1, 2C_3, 3\sigma\}$	Conjugated subgroups of the spatial symmetry group T_d.
4	$C_3 \wedge \mathcal{T}_s$	$\{1, 2C_3, 3\mathcal{T}_s\}$	This group has $\mathcal{T}_s$ instead of σ_d in C_{3v} and is isomorphic to C_{3v} as an abstract group.[8]

close to the limit of linearization (i.e., at any arbitrarily small perturbation ϵ). In the simplest case, RE are entirely defined by the finite symmetry $T_d \times \mathcal{T}$ of our system. They lie on the *critical orbits* of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ (they are isolated fixed points of this action). The position of these orbits is independent of the interaction potential (and thus of the particular Hamiltonian). We combine information about critical orbits on each of the factor spaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ in order to find all critical orbits on the total space. Considering Morse theory requirements, local symmetry, and local symplectic coordinates, we suggest possible stabilities of RE.

Sections 6–9 focus on the dynamical analysis of the reduced system; concrete applications are presented in sections 10 and 11. As soon as the interaction potential and the Hamiltonian

[4–9] See the footnotes to Table 3 on the following page.

Table 3

Classes of conjugated and invariant subgroups of the $T_d \times \mathcal{T}$ group. Part II.

	Class[4]	Structure[5]	Description and comments
	Subgroups of order 8		
1	$D_2 \times \mathcal{T}$	$\{\,1, C_2^x, C_2^y, C_2^z,$ $\mathcal{T}, \mathcal{T}_2^x, \mathcal{T}_2^y, \mathcal{T}_2^z\}$	Direct product of the D_2 subgroup of T_d and time reversal $\mathcal{T}$.
3	$C_{2v} \times \mathcal{T}$	$\{\,1, C_2^a,\, \sigma^{a_1}, \sigma^{a_2},$ $\mathcal{T}, \mathcal{T}_2^a, \mathcal{T}_s^{a_1}, \mathcal{T}_s^{a_2}\}$	Direct product of the C_{2v}^a subgroup of T_d and time reversal $a = (x, y, z)$.
3	D_{2d}	$\{1, C_2^a, \sigma^{a_{1,2}}, C_2^{b,c}, S_4^{\pm 1}\}$	Conjugated subgroups of the spatial symmetry group T_d.
3	$C_{2v} \wedge \mathcal{T}_2$	$\{1, C_2^a, \sigma^{a_{1,2}}, \mathcal{T}_2^{b,c}, \mathcal{T}_{\pm 4}^a\}$	Isomorphic to D_{2d} as an abstract group[6,7,9]; two conjugated C_2 rotations and two conjugated S_4 operations of D_{2d} are replaced for their products with $\mathcal{T}$.
3	$D_2 \wedge \mathcal{T}_s$	$\{1, C_2^a, \mathcal{T}_s^{a_{1,2}}, C_2^{b,c}, \mathcal{T}_{\pm 4}^a\}$	Isomorphic to D_{2d} as an abstract group; has two conjugated σ reflections and two conjugated S_4 operations[7,8,9] of D_{2d} replaced for their products with $\mathcal{T}$.
3	$S_4 \times \mathcal{T}$	$\{\,1, S_4, C_2, S_4^{-1},$ $\mathcal{T}, \mathcal{T}_4,\, \mathcal{T}_2, \mathcal{T}_{-4}\}$	Direct product of the cyclic subgroup S_4^a of T_d and time reversal.[9]
3	$S_4 \wedge \mathcal{T}_2$ $(S_4 \wedge \mathcal{T}_s)$	$\{1, C_2^a, \mathcal{T}_s^{a_{1,2}}, \mathcal{T}_2^{b,c}, S_4^{\pm 1}\}$	Isomorphic to D_{2d} as an abstract group; has two conjugated σ_d reflections and two conjugated C_2 rotations[6,7,8,9] of D_{2d} replaced for their products with time reversal $\mathcal{T}$.
	Subgroups of order 12		
1	T		Rotational subgroup of the T_d group.
4	$C_{3v} \times \mathcal{T}$		Direct product of a C_{3v} subgroup of T_d and time reversal.
	Subgroups of order 16		
3	$D_{2d} \times \mathcal{T}$		Direct product of one of the three D_{2d}^a subgroups of T_d and time reversal $\mathcal{T}$.
	Subgroups of index two (order 24)		
1	$T \times \mathcal{T}$		Direct product of T and time reversal, also denoted as $T \times Z_2$.
1	T_d		Tetrahedral group, isomorphic to the permutation group π_4 as an abstract group.
1	$T \wedge \mathcal{T}_s$		Another realization of π_4 obtained from T_d by replacing all improper rotations, namely six σ^α and six S_4 operations, for their products with time reversal $\mathcal{T}$.
	Complete group (order 48)		
1	$T_d \times \mathcal{T}$		Direct product of T_d and time reversal group $\mathcal{T}$.

[4] The leftmost column gives the number of conjugated subgroups in the class.

[5] When all operations in the group correspond to the same rotation axis a, we do not specify the choice of the axis and omit index a.

[6] The operation $\mathcal{T}_2^a = C_2^a \circ \mathcal{T}$ is rotation by π around axes a and time reversal; by convention $a = (x, y, z)$ is one of the axes S_4.

[7] In the T_d group, reflection planes σ^{a_1} and σ^{a_2} intersect on axis C_2^a, where by convention a is one of (x, y, z); in the O_h group these planes are called σ_d.

[8] The operation $\mathcal{T}_s = \sigma \circ \mathcal{T}$ is reflection in one of the six planes σ_d and time reversal; in particular, $\mathcal{T}_s^{a_{1,2}} = \mathcal{T} \circ \sigma^{a_{1,2}}$.

[9] The operations $\mathcal{T}_{\pm 4}^a = \mathcal{T}(S_4^a)^{\pm 1} = \mathcal{T} \circ (S_4^a)^{\pm 1}$, where $a = (x, y, z)$, are operations S_4 or S_4^{-1} combined with time reversal $\mathcal{T}$.

are introduced explicitly, the value of H_{eff} at critical orbits is found. This gives analytic expressions for the energy of RE as a function of actions N_e, N_f, and J. Using these energies, we characterize the multiplet of quantum states with quantum numbers N_e, N_f, and J. We can also use the quantum analogue of H_{eff} in order to compute energies of individual states. Using the concrete Hamiltonian, we can check the Morse indexes of all known RE and find, if necessary, additional RE which do not lie on critical orbits but are nevertheless required by Morse theory conditions.

3. Finite symmetry group of the system. We briefly review the structure of the tetrahedral group T_d and its extension $T_d \times \mathcal{T}$. This group and its subgroups can be considered as magnetic (or color) crystallographic symmetry groups; see Chapter 2 of [67]. Notation for such groups is not commonly established. Below we explain our conventions and describe the abstract group structure of $T_d \times \mathcal{T}$ given by its subgroup lattice (see Tables 2 and 3). We distinguish only the nonconjugate subgroups of $T_d \times \mathcal{T}$ and study certain sublattices corresponding to reduced or partial symmetry groups. This information is vital for understanding the stratification of different reduced phase spaces by the action of $T_d \times \mathcal{T}$ and, in particular, for finding fixed points and invariant subspaces of this action.

3.1. Time reversal symmetry $\mathcal{T}$. Momentum reversal symmetry $\mathcal{T}$ is a nonsymplectic symmetry operation defined for the original physical 3-space coordinates and conjugate momenta as

$$(q, p) \to (q, -p).$$

We denote this operation as $\mathcal{T}$, or simply as Z_2, and imply that its action in each particular context is either known or should be specified. We will distinguish the two types of behavior (two representations of Z_2) with regard to momentum reversal by “parity” indexes g (*gerade*) and u (*ungerade*), respectively.

The symmetry operation $\mathcal{T}$ is also sometimes called *time reversal*. We like to make clear that our operation $\mathcal{T}$ acts only on the phase space variables (q, p) and does not involve time t. This implies that the action of $\mathcal{T}$ on the extended phase space is

$$\mathcal{T} : (q, p, t) \to (q, -p, t).$$

It can be seen that, even when the Hamiltonian of the system is invariant with regard to such operation $\mathcal{T}$, the corresponding equations of motion are not. In fact, this is due to the fact that the action of $\mathcal{T}$ on (q, p) is antisymplectic. Operation $\mathcal{T}$ is an example of reversing symmetries. Another commonly used definition of time reversal extends nontrivially our operation $\mathcal{T}$ to time t:

$$\mathcal{T}_t : (q, p, t) \to (q, -p, -t).$$

This operation preserves the flow of the system with $\mathcal{T}$-invariant Hamiltonian function. In quantum mechanics, operation $\mathcal{T}$ changes the signs of all commutator relations while $\mathcal{T}_t$ preserves these signs.

Of course, one can use either $\mathcal{T}$ or $\mathcal{T}_t$ for the analysis, as long as one understands their action. For example, action of $\mathcal{T}$ and $\mathcal{T}_t$ on equilibria is the same, while their action on the trajectories $\gamma : t \to (q, p)$ is different. Trajectories $\mathcal{T}(\gamma)$ and $\mathcal{T}_t(\gamma)$ coincide in the phase space but have different direction. Reversing direction is the result of $t \to -t$, which cannot be represented as a geometric transformation of the phase space. Since in most cases we do not work with extended phase spaces, we prefer to define $\mathcal{T}$ as just another transformation of the phase space with coordinates (q, p).

The action of $\mathcal{T} \sim Z_2$ on vibrational normal mode coordinates and conjugate momenta is, of course, the same as on the 3-space q's and p's, and as we go to the complex variables

$$z = q - ip, \quad \bar{z} = q + ip,$$

this action becomes

$$(z_1, z_2, z_3, z_4, z_5) \to (\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4, \bar{z}_5). \tag{3.1a}$$

Note that this operation differs from plain "complex conjugation" as shown by

$$z_1 + iz_2 \to \bar{z}_1 + i\bar{z}_2. \tag{3.1b}$$

Molecular angular momentum components (j_1, j_2, j_3) are not invariant with respect to Z_2:

$$(j_1, j_2, j_3) \to (-j_1, -j_2, -j_3). \tag{3.1c}$$

This property of (j_1, j_2, j_3) follows, of course, from the explicit Wilson–Howard definition of rotational angular momenta in terms of particle coordinates and momenta. At the same time, we can simply note that time reversal changes the direction of classical rotation and therefore changes signs of (j_1, j_2, j_3).

3.2. Spatial finite symmetry T_d. The spatial symmetry group of the A_4 molecule is the point group of its tetrahedral equilibrium configuration T_d. This group, and cubic groups O and O_h (and to a lesser extent, T and T_h) are well known to molecular physicists and crystallographers. It is generated by the three-fold rotation C_3 and the four-fold inversion rotation S_4. The latter can be realized as $C_4 \circ C_i$, a rotation C_4 by angle $\pi/2$ followed by a 3-space inversion C_i, or alternately as rotation by angle $-\pi/2$ followed by reflection in the plane orthogonal to the rotation axis. A particular realization of T_d is given in Table 4 and is illustrated in Figure 1. We will use the three symmetry operations in Table 4 for explicit demonstrations later in the paper. Conventionally, axes x, y, and z are chosen as S_4 axes. Three operations S_4 and three inverse operations S_4^{-1} form a class of six conjugate elements. Three operations $C_2 = S_4^2$, which rotate by π about the same axes, form a separate class. Operations C_3, which rotate by $\pm 2\pi/3$ about four diagonal axes, such as axis $[1, 1, 1]$ in Table 4, form one class of eight elements. Finally, there is a class of six reflection planes denoted σ or C_s; each element C_s can be considered as a combination $C_2 \circ C_i$, where axis C_2 is orthogonal to the reflection plane.

3.3. Full finite symmetry group $T_d \times \mathcal{T}$. The total symmetry group of our system is the tetrahedral group T_d extended to include the time reversal operation $\mathcal{T}$. We exploit the isomorphism $T_d \times \mathcal{T} \sim O_h \sim O \times \mathcal{T}$ to explain the notation for symmetry operations and different subgroups of $T_d \times \mathcal{T}$. Table 5 summarizes correspondence of notation for the symmetry operations of the three groups.

Table 4

Matrix representations of basic operations of the T_d group illustrated in Figure 1.

	A_1	A_2	F_2	E	F_1
	z_a		(z_1, z_2, z_3)	(z_4, z_5)	(j_1, j_2, j_3)
$C_3^{[111]}$	1	1	$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$	$\frac{1}{2}\begin{pmatrix} -1 & -\sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
S_4^z	1	−1	$\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	$\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
C_s^{xy}	1	−1	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$	$-\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

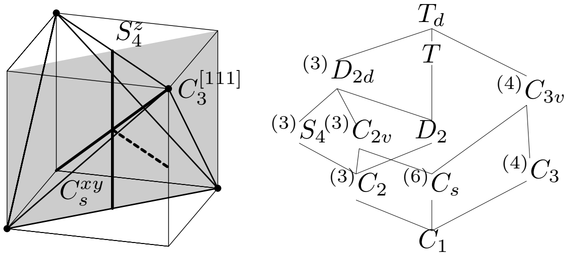


Figure 1. *Basic operations of the T_d point group (left). The symmetry axis C_2 (dashed line) of the O_h and O groups is orthogonal to the C_s reflection plane (shadowed) of T_d. This C_2 should not be confused with axis $C_2 = S_4^2$, which has the same orientation as axis S_4. Right: lattice of conjugate subgroups of the T_d group. Left: superscripts give the number of conjugate subgroups in each class.*

The lattices of conjugate subgroups of the O_h and $T_d \times \mathcal{T}$ groups are shown in Figures 2 and 3, respectively, in order to compare the Schönflis notation for the classes of O_h [67] to our notation of the $T_d \times \mathcal{T}$ classes. The 33 classes of conjugate subgroups of $T_d \times \mathcal{T}$ are arranged according to their order and are further described in Tables 2 and 3. Left superscripts in Figures 2 and 3 indicate, where necessary, the number of conjugate subgroups in the class. Invariant subgroups are unique in their class which needs, therefore, no such superscripts. Subgroups of O_h, which are distinguished by primes, C_{2v}, C'_{2v}, and C''_{2v}, D_{2d} and D'_{2d}, D_{2h} and D'_{2h}, D_2 and D'_2, C_{2h} and C'_{2h}, C_2 and C'_2, C_s and C'_s, are nonconjugate in O_h but become conjugate in the larger group SO(3). Such notation is less informative in comparison with the $T_d \times \mathcal{T}$ notation for the corresponding nonconjugate subgroups, which highlights the differences between the subgroups explicitly.

3.4. Sublattices corresponding to different images of $T_d \times \mathcal{T}$ and broken symmetries.

The action of the symmetry group $T_d \times \mathcal{T}$ on the vibrational E-mode polyad space $\mathbb{C}P^1$ is not effective; the invariant subgroup D_2 forms the kernel, and the image $(T_d \times \mathcal{T})/D_2$ is the group isomorphic (as an abstract group) to $C_{3v} \times \mathcal{T}$ or D_{3h}, see Table 6. We compare the subgroup lattice of $(T_d \times \mathcal{T})/D_2$ in Figure 4, right, to the equivalent lattice of D_{3h} (Figure 4, left) in order to better explain the action of $T_d \times \mathcal{T}$ on $\mathbb{C}P^1$. When characterizing the subgroups of $(T_d \times \mathcal{T})/D_2$ we take into account their extension by the kernel D_2 in order to preserve the

Table 5

Correspondence between classes of conjugate elements of the groups O_h, $T_d \times \mathcal{T}$, and $O \times \mathcal{T}$.

O_h	$T_d \times \mathcal{T}$	$O \times \mathcal{T}$	O_h	$T_d \times \mathcal{T}$	$O \times \mathcal{T}$
1	1	1	C_i	$\mathcal{T}$	$\mathcal{T}$
$8C_3$	$8C_3$	$8C_3$	$8S_6$	$8(\mathcal{T}C_3)$	$8(\mathcal{T}C_3)$
$3C_2$	$3C_2$	$3C_2$	$3\sigma_h$	$3(\mathcal{T}C_2)$	$3(\mathcal{T}C_2)$
$6C_2'$	$6\sigma_d$	$6C_2'$	$6\sigma_d$	$6(\mathcal{T}\sigma_d)$	$6(\mathcal{T}C_2')$
$6C_4$	$6S_4$	$6C_4$	$6S_4$	$6(\mathcal{T}S_4)$	$6(\mathcal{T}C_4)$

Table 6

Homomorphism $T_d \times \mathcal{T} \to D_{3h}$, which defines the group of symmetry transformations of the vibrational reduced polyad phase space $\mathbb{C}P^1$ of the doubly degenerate mode E. Each element in D_{3h} is an image of four elements of $T_d \times \mathcal{T}$. In particular, the identity in D_{3h} is the image of the invariant subgroup D_2 of $T_d \times \mathcal{T}$.

$\{T_d \times \mathcal{T}\}$	$\to$	D_{3h}	$\{T_d \times \mathcal{T}\}$	$\to$	D_{3h}
$\{1, 3C_2\}$	$\to$	1	$\{\mathcal{T}, 3(\mathcal{T}C_2)\}$	$\to$	σ_h
$\{8C_3\}$	$\to$	$2C_3$	$\{8(\mathcal{T}C_3)\}$	$\to$	$2S_3$
$\{6S_4, 6\sigma_d\}$	$\to$	$3C_2$	$\{6(\mathcal{T}S_4), 6(\mathcal{T}\sigma_d)\}$	$\to$	$3\sigma_v$

relation to the $T_d \times \mathcal{T}$ action on the subspaces $\mathbb{C}P^2$ and $\mathbb{S}^2$ (where $T_d \times \mathcal{T}$ acts effectively).

The $T_d \times \mathcal{T}$ group has a number of subgroups whose action on classical phase spaces $\mathbb{C}P^2$ and $\mathbb{S}^2$ was described earlier in [17, 70, 18]; action of the $D_2 \times \mathcal{T}$ group on $\mathbb{C}P^2$ was studied in detail in [70]. The Schönflis notation for spatial finite groups used in these studies can be misleading if the actual spatial-temporal symmetry operations are not defined explicitly. In order to compare our present work to [70] we give in Figure 5 the correspondence of the subgroup lattice of $D_2 \times \mathcal{T}$ to that of the group D_{2h}. These two groups realize the same abstract Abelian group of order 8, and all their subgroups are invariant. As shown in Figure 5, certain invariant subgroups of $D_2 \times \mathcal{T}$ (or D_{2h}) can be assembled in sets of three according to the axes C_2^a, $a = \{x, y, z\}$, of the D_2 group. These subgroups become conjugate when lifted to the higher symmetry group $T_d \times \mathcal{T}$ (or O_h).

4. Group action. When a group element acts on the point x on the space $\mathcal{P}$, it can map x either to a different point x' on $\mathcal{P}$ or to itself. In the latter case the group element belongs to the local symmetry group or *stabilizer* of x. The set of points obtained from x by applying all group elements is called an *orbit*. Orbits of a finite group G are, obviously, finite sets, and the maximum number of points in an orbit of the G action equals the number of group elements or the order $[G]$ of the group. If the stabilizer G_x of the point x is nontrivial, then the number of points in the orbit equals $[G]/[G_x]$. In particular, if x is a fixed point of the group action, it forms a one-point orbit.

Out of all the orbits of the action of the group G on the space $\mathcal{P}$, we distinguish *critical orbits* [75, 76]. The stabilizer of a point x on the critical orbit differs from that of any point in a sufficiently small open neighborhood of x; i.e., points on critical orbits are isolated. As a consequence, these points must be stationary or critical points of any G-invariant function $f(x)$ on $\mathcal{P}$. When $\mathcal{P}$ is a reduced phase space and $f(x)$ is a reduced Hamiltonian H_{eff}, these points correspond to RE of the initial system. This makes finding critical orbits the primary purpose of our group action study [48, 49, 50]. In general, we also look for invariant subspaces

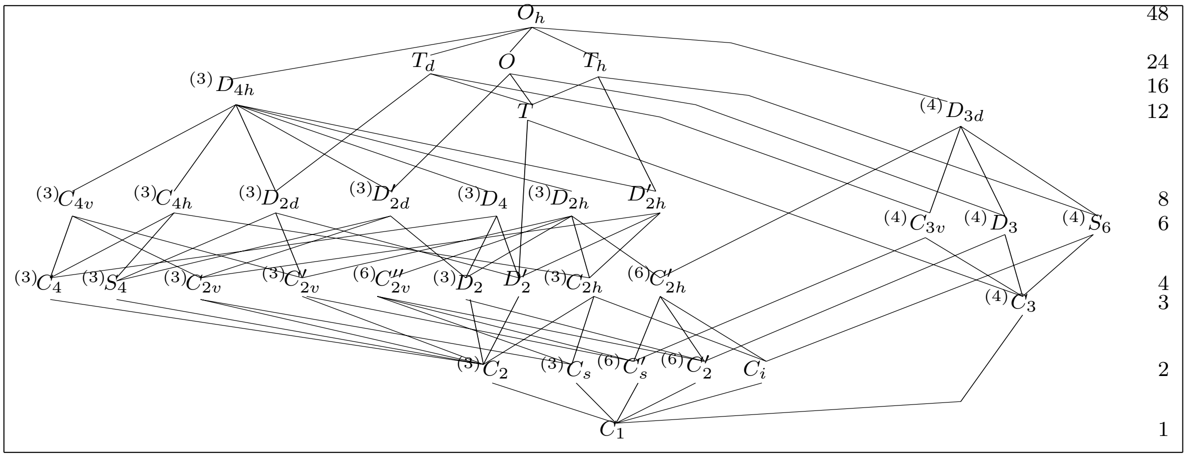


Figure 2. *Lattice of conjugate subgroups of the O_h group. The order of all subgroups on the same row is indicated on the right of the graph.*

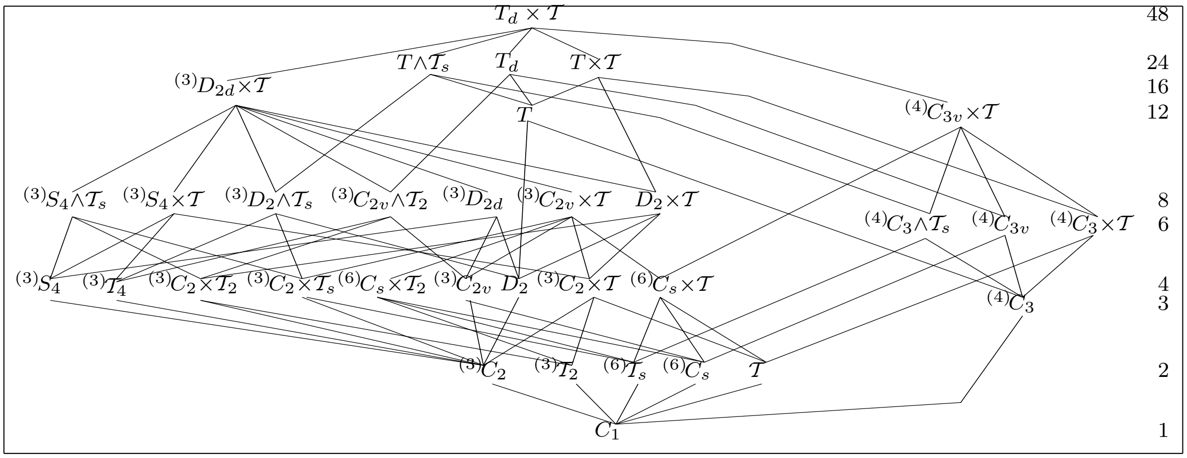

$T_d \times \mathcal{T}$
48
$\mathcal{T}\wedge\mathcal{T}_s$
T_d
$\mathcal{T}\times\mathcal{T}$
24
16
$^{(3)}D_{2d}\times\mathcal{T}$
$\mathcal{T}$
12
$^{(4)}C_{3v}\times\mathcal{T}$
$^{(3)}S_4\wedge\mathcal{T}_s$
$^{(3)}S_4\times\mathcal{T}$
$^{(3)}D_2\wedge\mathcal{T}_s$
$^{(3)}C_{2v}\wedge\mathcal{T}_2$
$^{(3)}D_{2d}$
$^{(3)}C_{2v}\times\mathcal{T}$
$D_2\times\mathcal{T}$
8
$^{(4)}C_3\wedge\mathcal{T}_s$
$^{(4)}C_{3v}$
$^{(4)}C_3\times\mathcal{T}$
6
$^{(3)}S_4$
$^{(3)}\mathcal{T}_4$
$^{(3)}C_2\times\mathcal{T}_2$
$^{(3)}C_2\times\mathcal{T}_s$
$^{(6)}C_s\times\mathcal{T}_2$
$^{(3)}C_{2v}$
D_2
$^{(3)}C_2\times\mathcal{T}$
$^{(6)}C_s\times\mathcal{T}$
4
$^{(4)}C_3$
3
$^{(3)}C_2$
$^{(3)}\mathcal{T}_2$
$^{(6)}\mathcal{T}_s$
$^{(6)}C_s$
$\mathcal{T}$
2
C_1
1

Figure 3. *Lattice of conjugate subgroups of the $T_d \times \mathcal{T}$ group; cf. Figure 2. Shorthand notation $\mathcal{T}_2$, $\mathcal{T}_s$, and $\mathcal{T}_4$ is used for stabilizers $\mathcal{T}C_2$, $\mathcal{T}\sigma_d$, and $\mathcal{T}S_4$; the order of all subgroups on the same row is indicated on the right of the graph.*

of the reduced phase space $\mathcal{P}$—and especially for the invariant subspaces whose stabilizer is a purely spatial symmetry subgroup of G.

The symmetry group T_d was originally defined as a point group of transformations in the Euclidean 3-space $\mathbb{R}^3$ with coordinates (x, y, z), which transform in the same way as components of the F_2 mode (q_1, q_2, q_3). The action of T_d is subsequently extended symplectically on (p_1, p_2, p_3), which transform in the same way as (q_1, q_2, q_3). This defines the action on (z_1, z_2, z_3). At the same time, momentum reversal $\mathcal{T}$ or Z_2 is introduced as an antisymplectic symmetry.

The E-mode variables and rotational variables (j_1, j_2, j_3) transform according to the E and F_1 irreducible representations of the T_d group. The action of the symmetry group in these

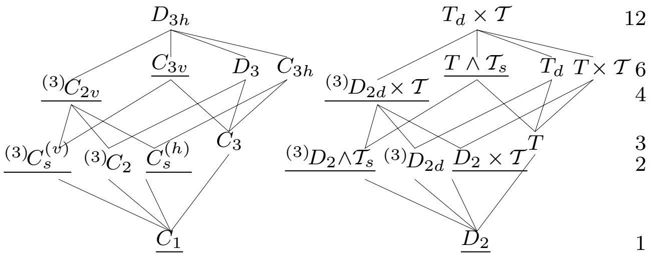


Figure 4. *Left: lattice of conjugate subgroups of the D_{3h} group. Underlined subgroups appear as stabilizers of the strata of the D_{3h} action on the $\mathbb{S}^2$ sphere. Right: part of the lattice of conjugate subgroups common to $T_d \times \mathcal{T}$ and $(T_d \times \mathcal{T})/D_2$. Underlined subgroups appear as stabilizers of the $T_d \times \mathcal{T}$ action on the E-mode polyad reduced phase space $\mathbb{C}P^1 \sim \mathbb{S}^2$.*

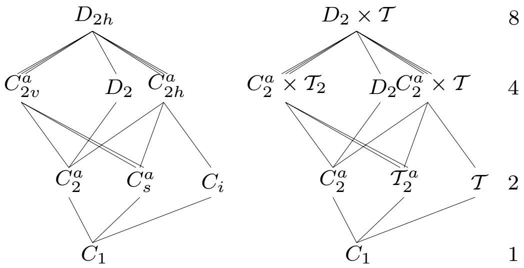


Figure 5. *Subgroup lattices of the D_{2h} group (left) and the $D_2 \times \mathcal{T}$ subgroup of $T_d \times \mathcal{T}$ (right); subgroups distinguished by superscript $a = \{x, y, z\}$ are conjugate in the higher groups O_h and $T_d \times \mathcal{T}$, respectively.*

representations is defined by the *image* of the initial symmetry group. To find the action of the symmetry group T_d and its extension $T_d \times \mathcal{T}$ on different components of the reduced phase space (the F_2-mode space $\mathbb{C}P^2$, the E-mode space $\mathbb{C}P^1$, and the rotational sphere $\mathbb{S}^2$), we first need to know the image of our symmetry group in the corresponding representations. We find the image of the group in the particular representation Γ by acting explicitly on the variables which realize Γ.

Group images and their actions can be very nontrivial even for finite symmetry groups and should be studied with care. Thus the action of spatial inversion on the reduced phase space of our system is equivalent to identity, and as a consequence, it suffices to consider pure rotations C_4 and C_2 of the O group instead of operations S_4 and C_s of the T_d group. We will also see that the image of T_d in the E representation is a smaller group C_{3v} and that its action on the E-mode reduced phase space $\mathbb{C}P^1$ is equivalent to that of a dihedral group D_3. We begin by explaining actions of $T_d \times \mathcal{T}$ on the reduced phase spaces $\mathbb{C}P^2$, $\mathbb{C}P^1$, and $\mathbb{S}^2$ and on the total reduced space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ with the action of the rotation group SO(2) (or C_∞) and its finite subgroups C_k, $k = 2, 3, 4, \dots$, of which C_2 is a special case, and by explaining the action of time reversal $\mathcal{T}$ (or Z_2). We use operations from Table 4 to illustrate group actions.

4.1. Rotational subsystem: Action of $T_d \times \mathcal{T}$ on $\mathbb{S}^2$.

Unlike components of polar vectors (x, y, z) and (q_1, q_2, q_3), the angular momenta (j_1, j_2, j_3) are invariant with regard to the 3-space inversion C_i (i.e., they do not change sign). The image of T_d in the F_1 representation realized by (j_1, j_2, j_3) is an isomorphic group O generated by pure rotations. (This group can

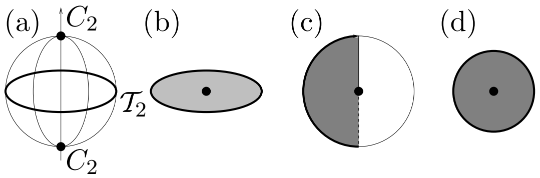


Figure 6. *Action of the $C_2 \times \mathcal{T}$ group on the rotational sphere $\mathbb{S}^2$ and construction of the corresponding orbit space.* (a) *Two points with stabilizer C_2 are shown as filled black circles; points with stabilizer $\mathcal{T}_2$ lie on the circle shown in bold.* (b) *The $\mathcal{T}_2$ symmetry is reduced; all points inside the disc represent two-point orbits.* (c)–(d) *The C_2 and $\mathcal{T}$ symmetries are reduced; the disc is cut and glued conewise. All points in the shaded interior in* (d) *represent generic four-point orbits with stabilizer C_1; the black circle corresponds to the two-point orbit with stabilizer C_2; two-point orbits with stabilizer $\mathcal{T}_2$ form the boundary, which is a one-dimensional stratum $\mathbb{S}^1$.*

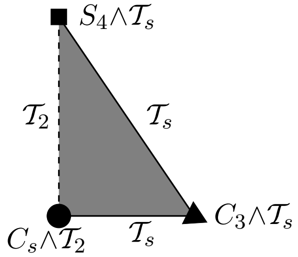


Figure 7. *Space of orbits of the $T_d \times \mathcal{T}$ symmetry group action on the rotational phase space $\mathbb{S}^2$. Critical orbits with stabilizers $S_4 \wedge \mathcal{T}_s$, $C_3 \wedge \mathcal{T}_s$, and $C_s \wedge \mathcal{T}_2$ (C_{4v}, C_{3v}, and C_{2v} in the O_h notation) are shown by the black square, triangle, and disc, respectively. One-dimensional strata with stabilizers $\mathcal{T}_s$ (C_s') and $\mathcal{T}_2$ (C_s) are shown by solid and dashed border lines. Generic C_1 orbits correspond to points in the shaded interior.*

be obtained from T_d if S_4 is replaced with $C_4 = S_4 \circ C_i$ and C_s with $C_2 = C_s \circ C_i$; see Table 5.) On the other hand, time reversal $\mathcal{T}$ changes the signs of all three components of the angular momentum, and therefore (j_1, j_2, j_3) realize the F_{1u} representation of $T_d \times \mathcal{T}$. The image of $T_d \times \mathcal{T}$ in this representation is the group $O \times \mathcal{T}$. The three isomorphic groups $T_d \times \mathcal{T}$, $O \times \mathcal{T}$, and O_h are realizations of the same abstract group. Correspondence of their subgroups and classes of conjugate elements is presented in Figures 2 and 3 and Table 5.

As discussed in section 2.4.1, equation $j^2 = \text{const}$ defines the reduced rotational phase space $\mathbb{S}^2$ as a sphere in the ambient space $\mathbb{R}^3$ with coordinates (j_1, j_2, j_3). The action of $O \times \mathcal{T}$ on this sphere is often represented in terms of the action of the O_h group of transformations of the $\mathbb{R}^3$ space (see Figure 7). The O_h notation, or even shorter O group notation, is used in practically all applications [18, 39, 32, 33, 34] and remains preferred (at least for the study of a purely rotational system) because geometric transformations in $\mathbb{R}^3$ are very commonly known. On the other hand, the $T_d \times \mathcal{T}$ notation properly reflects the actual symmetry of the system and is more adequate.

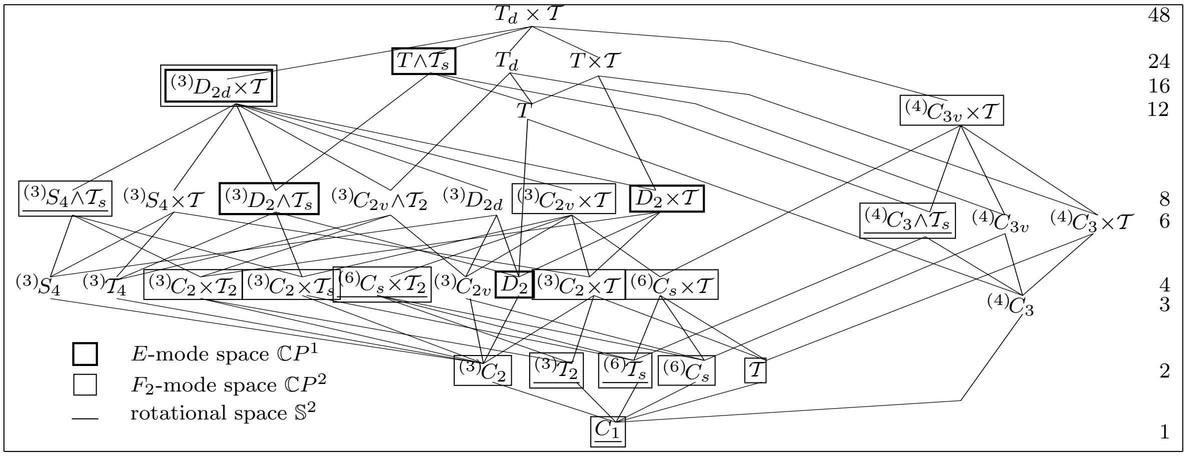
Any rotation C_k acting on $\mathbb{S}^2$ (as an element of the O_h group of transformations of the ambient space $\mathbb{R}^3$) has *two* fixed points, which are the two diametrically opposite points of $\mathbb{S}^2$ situated on the axis. The two points are mapped into each other by the $\mathcal{T}$ operation (which acts as inversion in $\mathbb{R}^3$) and belong to one orbit. This orbit is critical. Reversing operations $C_k \circ \mathcal{T}$ with $k > 2$ have no fixed points on $\mathbb{S}^2$. The operation $\mathcal{T}_2 = C_2 \circ \mathcal{T}$ acts in $\mathbb{R}^3$ like a symmetry plane σ orthogonal to the C_2 axis. The set of all points on $\mathbb{S}^2$ invariant with regard to $\mathcal{T}_2$ is a circle $\mathbb{S}^1$, which is the intersection of σ and $\mathbb{S}^2$.

As a simple example, consider the action of the $C_2 \times \mathcal{T}$ group on $\mathbb{S}^2$ illustrated in Figure 6

Table 7

Action of $T_d \times \mathcal{T}$ on the rotational sphere $\mathbb{S}^2$.

O_h stabilizer	C_{4v}	C_{3v}	C_{2v}	C_s	C'_s	C_1
$T_d \times \mathcal{T}$ stabilizer	$S_4 \wedge \mathcal{T}_s$	$C_3 \wedge \mathcal{T}_s$	$C_s \wedge \mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_s$	C_1
Points in orbit	6	8	12	24	24	48
Conjugate stabilizers	3	4	6	3	6	1



$T_d \times \mathcal{T}$
48
24
16
12
8
6
4
3
2
1
$^{(3)}D_{2d}{\times}\mathcal{T}$
$\mathcal{T}{\wedge}\mathcal{T}_s$
T_d
$\mathcal{T}{\times}\mathcal{T}$
$\mathcal{T}$
$^{(4)}C_{3v}{\times}\mathcal{T}$
$^{(3)}S_4{\wedge}\mathcal{T}_s$
$^{(3)}S_4{\times}\mathcal{T}$
$^{(3)}D_2{\wedge}\mathcal{T}_s$
$^{(3)}C_{2v}{\wedge}\mathcal{T}_2$
$^{(3)}D_{2d}$
$^{(3)}C_{2v}{\times}\mathcal{T}$
$D_2{\times}\mathcal{T}$
$^{(4)}C_3{\wedge}\mathcal{T}_s$
$^{(4)}C_{3v}$
$^{(4)}C_3{\times}\mathcal{T}$
$^{(3)}S_4$
$^{(3)}\mathcal{T}_4$
$^{(3)}C_2{\times}\mathcal{T}_2$
$^{(3)}C_2{\times}\mathcal{T}_s$
$^{(6)}C_s{\times}\mathcal{T}_2$
$^{(3)}C_{2v}$
D_2
$^{(3)}C_2{\times}\mathcal{T}$
$^{(6)}C_s{\times}\mathcal{T}$
$^{(4)}C_3$
$^{(3)}C_2$
$^{(3)}\mathcal{T}_2$
$^{(6)}\mathcal{T}_s$
$^{(6)}C_s$
$\mathcal{T}$
C_1
E-mode space $\mathbb{C}P^1$
F_2-mode space $\mathbb{C}P^2$
rotational space $\mathbb{S}^2$

Figure 8. *Lattice of conjugate subgroups of the $T_d \times \mathcal{T}$ group. Subgroups that appear as stabilizers on the reduced phase spaces $\mathbb{S}^2$, $\mathbb{C}P^1$, and $\mathbb{C}P^2$ are underlined, bold framed, and framed, respectively. The order of all subgroups on the same row is indicated on the right of the graph; cf. Figure 3.*

(with the C_2 axis along the z axis). There are three types of orbits: a two-point orbit with stabilizer C_2, a one-dimensional stratum of two-point orbits with stabilizer $\mathcal{T}_2$, and a two-dimensional stratum with trivial stabilizer. The space of orbits is a punctured closed disc shown in Figure 6, right.

The action of $T_d \times \mathcal{T}$ on $\mathbb{S}^2$ is described in Table 7 and the space of orbits is shown in Figure 7. Out of 33 classes of the conjugate subgroups of this group, six appear as stabilizers (see Figure 8). There are 26 fixed points on $\mathbb{S}^2$, which form three critical orbits with stabilizers $S_4 \wedge \mathcal{T}_s \sim C_{4v}$, $C_3 \wedge \mathcal{T}_s \sim C_{3v}$, $C_s \wedge \mathcal{T}_2 \sim C_{2v}$. Note that, as in the case of any C_k action, each specific stabilizer in the class of conjugate stabilizers corresponds to two different points in the orbit.

4.2. *E*-mode vibrational subsystem: Action of $T_d \times \mathcal{T}$ on $\mathbb{C}P^1 \sim \mathbb{S}^2$. We find the image of the spatial symmetry group T_d in the E representation by considering the action of T_d on the plane $\mathbb{R}^2$ with coordinates (q_4, q_5) or on a complex plane with coordinates (z_4, z_5). From Table 4 we can see that operation $C_2^z = (S_4^z)^2$ acts trivially on this plane, operation C_3 rotates by $2\pi/3$ about the origin, while operations S_4^z and C_s^{xy} have the *same* action on $\mathbb{R}^2$: they reflect with respect to the axis $\{q_5 = 0\}$ passing through the origin. It follows that the image of T_d is a group D_3 (or C_{3v}).

The action of the full symmetry group $T_d \times \mathcal{T}$ on the reduced vibrational phase space $\mathbb{C}P^1$ is equivalent to that of $C_{3v} \times \mathcal{T}$. The kernel of the homomorphism $T_d \times \mathcal{T} \to C_{3v} \times \mathcal{T}$

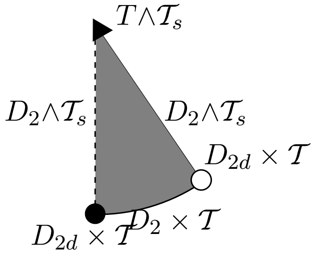


Figure 9. *Space of orbits of the $T_d\times\mathcal{T}$ symmetry group action on the E-mode reduced phase space $\mathbb{C}P^1\sim\mathbb{S}^2$. Triangle and circles mark critical orbits with stabilizers $T\wedge\mathcal{T}_s$ and $D_{2d}\times\mathcal{T}$ (C_{3v} and C_{2v} in the D_{3h} notation), respectively. One-dimensional strata with stabilizers $D_2\wedge\mathcal{T}_s$ and $D_2\times\mathcal{T}$ (σ_h) form the boundary of the variety, while generic D_2 orbits correspond to points in the interior.*

Table 8

Action of $T_d\times\mathcal{T}$ on the polyad phase space $\mathbb{C}P^1$ of the E mode; notation is explained in Table 6.

D_{3h} stabilizer	C_{3v}	C_{2v}	$C_s^{(v)}$	$C_s^{(h)}$	C_1
($T_d\times\mathcal{T}$) stabilizer	$T\wedge\mathcal{T}_s$	$D_{2d}\times\mathcal{T}$	$D_2\wedge\mathcal{T}_s$	$D_2\times\mathcal{T}$	D_2
Number of					
points in orbit	2	3	6	6	12
conj. stabilizers	1	3	3	1	1
orbits	1	2	∞	∞	∞^2

is D_2, i.e., the order four invariant subgroup of $T_d\times\mathcal{T}$ described in Table 2. The action of $C_{3v}\times\mathcal{T}$ on the E-mode vibrational phase sphere $\mathbb{S}^2\sim\mathbb{C}P^1$ can be better visualized as the natural action of the point group D_{3h} of transformations of the Euclidean 3-space $\mathbb{R}^3$ on a sphere embedded in this space. The correspondence between the D_{3h} notation and symmetry operations of $T_d\times\mathcal{T}$ is given in Figure 4 and Table 6.

All strata of the $T_d\times\mathcal{T}$ action on the E-mode space $\mathbb{C}P^1\sim\mathbb{S}^2$ are described in Figure 9 and Table 8. The D_{3h} analogy makes understanding this stratification straightforward. We can essentially use the approach in section 4.1 and, of course, earlier results for classical C_{3v} symmetric rotational systems (see section 1), such as a triatomic molecule with equilateral equilibrium configuration. For example, any rotation in D_{3h} has two fixed points on $\mathbb{S}^2$; the particular C_3 rotation in Table 4 has fixed points with coordinates $v=(0,\pm1,0)$ or equally $(z_4,z_5)=(1,\mp i)$, which form one two-point critical orbit with stabilizer $T\wedge\mathcal{T}_s$. The stabilizer of the two other critical orbits is $D_{2d}\times\mathcal{T}$. Note that the points in the $T\wedge\mathcal{T}_s$ orbit have exactly the same stabilizer (because $T\wedge\mathcal{T}_s$ is an invariant subgroup), whereas the three points in each $D_{2d}\times\mathcal{T}$ orbit have different conjugate stabilizers.

Dynamical variables of the reduced E-mode system and local canonical coordinates near points on critical orbits on $\mathbb{C}P^1$ can be classified using irreducible representations of both $T_d\times\mathcal{T}$ and D_{3h}. Thus, vibrational angular momenta $\{v_1,v_2,v_3\}$ introduced in section 2.4.2 span the natural reducible representation $A_{2u}\oplus E_g$ of the $T_d\times\mathcal{T}$ group (see section 7), which corresponds to $A_2''\oplus E'$ of the D_{3h} group. Table 9 gives the relation between the irreducible representations of the $T_d\times\mathcal{T}$ group and its subgroup $(T_d\wedge\mathcal{T})/D_2$. This correspondence should be taken into account in order to study vibration-rotation dynamics on $\mathbb{C}P^1\times\mathbb{S}^2$, where the group $T_d\times\mathcal{T}$ acts as D_{3h} and O_h on the vibrational subspace $\mathbb{C}P^1$ and rotational subspace $\mathbb{S}^2$. We should warn the reader that D_{3h} is *not* a subgroup of O_h. We simply use standard

Table 9

Reduction of irreducible representations of $T_d \times \mathcal{T}$ ($\sim O_h$) into those of $(T_d \times \mathcal{T})/D_2$ ($\sim D_{3h}$). The correspondence is written initially for O_h and its D_{3d} subgroup and is further extended using the isomorphism $D_{3d} \leftrightarrow D_{3h}$.

$T_d \times \mathcal{T} \to$			$(T_d \times \mathcal{T})/D_2$	$T_d \times \mathcal{T} \to$			$(T_d \times \mathcal{T})/D_2$
$\sim O_h$	$\to$	D_{3d}	$\sim D_{3h}$	$\sim O_h$	$\to$	D_{3d}	$\sim D_{3h}$
A_{1g}	$\to$	A_{1g}	A_1'	A_{1u}	$\to$	A_{1u}	A_1''
A_{2g}	$\to$	A_{2g}	A_2'	A_{2u}	$\to$	A_{2u}	A_2''
E_g	$\to$	E_g	E'	E_u	$\to$	E_u	E''
F_{1g}	$\to$	$A_{1g} \oplus E_g$	$A_1' \oplus E'$	F_{1u}	$\to$	$A_{1u} \oplus E_u$	$A_1'' \oplus E''$
F_{2g}	$\to$	$A_{2g} \oplus E_g$	$A_2' \oplus E'$	F_{2u}	$\to$	$A_{2u} \oplus E_u$	$A_2'' \oplus E''$

notation for irreducible representations of O_h and D_{3h}.

4.3. F_2-mode vibrational subsystem: Action on $\mathbb{C}P^2$. Action of several different point symmetry groups on the $\mathbb{C}P^2$ space was studied by Zhilinskií [17]. An example of the extension to a larger group including the time reversal $\mathcal{T}$ was given later in [70]. We summarize the results of [17] for the symmetry groups O and T_d and then take the $\mathcal{T}$ element into account.

The action of T_d on the 3-space with coordinates (q_1, q_2, q_3), which span the irreducible representation F_2, is effective, and the image corresponds to the whole group. Of course, the same holds for (p_1, p_2, p_3) and (z_1, z_2, z_3). In order to verify that the action of $T_d \times \mathcal{T}$ on the $\mathbb{C}P^2$ space is also effective [17], we can consider the representation realized by $(z_1, z_2, z_3, \bar{z}_1, \bar{z}_2, \bar{z}_3)$, act by operations in $T_d \times \mathcal{T}$ (use (3.1a) and Table 4), and take the $\mathbb{C}P^2$ restrictions (section 2.4.3) into account. In particular, due to the common phase equivalence, points (z_1, z_2, z_3) and $(-z_1, -z_2, -z_3)$ are the same on $\mathbb{C}P^2$; i.e., the image of the 3-space inversion is identity. It follows that the images of the T_d, O, and O_h point symmetry groups are the *same* (up to an isomorphism between the stabilizers), and we can simply use the O group whose elements are proper rotations.

For any rotation axis C_k, we should consider a point in $\mathbb{C}P^2$ lying on the axis and a subspace orthogonal to the axis. The former is obviously an isolated fixed point, while the latter is a $\mathbb{C}P^1 \sim \mathbb{S}^2$ subspace of $\mathbb{C}P^2$, which contains other C_k symmetric points. As an example, take rotation about axis z_1 that can be most easily understood in the coordinates

$$z_1, \quad \zeta = \frac{1}{\sqrt{2}}(z_2 + i z_3), \qquad \zeta' = \frac{1}{\sqrt{2}}(z_2 - i z_3),$$

subject to the same restriction

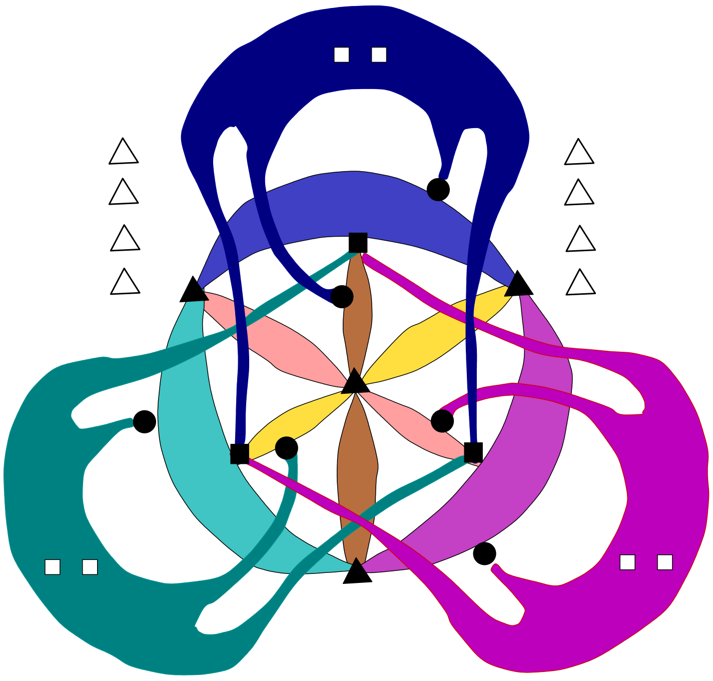
$$|z_1|^2 + |\zeta|^2 + |\zeta'|^2 = 1$$

and common phase identification

$$(z_1, \zeta, \zeta') \equiv (z_1 e^{i\phi}, \zeta e^{i\phi}, \zeta' e^{i\phi})$$

as the initial (z_1, z_2, z_3) in section 2.4.3. When we rotate about z_1 by angle $\varphi \neq 0$ so that

$$(z_1, \zeta, \zeta') \to (z_1, \zeta e^{i\varphi}, \zeta' e^{-i\varphi}),$$



fixed point	stabilizer in $T_d \times \mathcal{T}$
■	$D_{2d} \times \mathcal{T}$
▲	$C_{3v} \times \mathcal{T}$
□	$S_4 \wedge \mathcal{T}_s$
△	$C_3 \wedge \mathcal{T}_s$
•	$C_{2v} \times \mathcal{T}$

Figure 10. *Orbits of the O (and $T_d \times \mathcal{T}$) group action on the complex projective space $\mathbb{C}P^2$ according to Zhilinskií* [17]. *Colored areas represent nine C_2-invariant spheres, which belong to the classes of three and six spheres (according to their stabilizers).*

the fixed point coordinates should satisfy equations

$$(z_1, \zeta e^{i\varphi}, \zeta' e^{-i\varphi}) = (z_1 e^{i\phi}, \zeta e^{i\phi}, \zeta' e^{i\phi}).$$

For all $\varphi \neq \pi$ (and in particular for all $\varphi = 2\pi/k$ with $k > 2$) we have three isolated solutions

$$A : \zeta = \zeta' = 0, \quad B : z_1 = \zeta = 0, \quad \text{and} \quad C : z_1 = \zeta' = 0;$$

i.e., two of the three coordinates should vanish. When $\varphi = \pi$ (rotation C_2) our equations become

$$(z_1, -\zeta, -\zeta') = (z_1 e^{i\phi}, \zeta e^{i\phi}, \zeta' e^{i\phi}).$$

We should take the point $(1, 0, 0)$ and the whole $\mathbb{C}P^1$ subspace with $z_1 = 0$. (Set $\phi = \pi$ to show explicitly that $(0, z_2, z_3)$ and $(0, -z_2, -z_3)$ is the same point of this C_2-invariant subspace.)

The action of the entire symmetry group O on the reduced phase space $\mathbb{C}P^2$ is obtained if we apply the above principle to every rotation in O. The groups T_d and O have the same action on $\mathbb{C}P^2$. The action of $T_d \times \mathcal{T}$ then can be found as an extension by adding the $\mathcal{T}$ element. Results are summarized in Figure 10 reproduced from [17] and in Tables 10–12. Action of O on $\mathbb{C}P^2$ has five critical orbits (five zero-dimensional strata) characterized in Table 10. It is important to notice that each of the three points on the D_3 orbit correspond to a different (but conjugate) stabilizer; the same is true for the four points on the D_4 orbit. In

Table 10

Zero-dimensional strata of the $\mathbb{C}P^2$ space under the action of the image of the $T_d \times \mathcal{T}$ (or $O \times \mathcal{T}$ or $O_h \times \mathcal{T}$) group in the representation spanned by bilinear combinations of vibrational coordinates q and conjugate momenta p, which transform according to the triply degenerate representation F_2 of point symmetry group T_d, O, or O_h.

Orbit stabilizer[10]		Coordinates[11]	Point stabilizer[10]	
T_d	O or O_h	on $\mathbb{C}P^2$	O or O_h	T_d
$D_{2d} \times \mathcal{T}$	$D_4 \times \mathcal{T}$	$(1,0,0)$	$D_4^{(x)} \times \mathcal{T}$	$D_{2d}^{(x)} \times \mathcal{T}$
		$(0,1,0)$	$D_4^{(y)} \times \mathcal{T}$	$D_{2d}^{(y)} \times \mathcal{T}$
		$(0,0,1)$	$D_4^{(z)} \times \mathcal{T}$	$D_{2d}^{(z)} \times \mathcal{T}$
$C_{3v} \times \mathcal{T}$	$D_3 \times \mathcal{T}$	$(1,1,1)$	$D_3^{(c)} \times \mathcal{T}$	$C_{3v}^{(c)} \times \mathcal{T}$
		$(1,-1,-1)$	$D_3^{(d)} \times \mathcal{T}$	$C_{3v}^{(d)} \times \mathcal{T}$
		$(1,-1,1)$	$D_3^{(b)} \times \mathcal{T}$	$C_{3v}^{(b)} \times \mathcal{T}$
		$(1,1,-1)$	$D_3^{(a)} \times \mathcal{T}$	$C_{3v}^{(a)} \times \mathcal{T}$
$S_4 \wedge \mathcal{T}_2$	$C_4 \wedge \mathcal{T}_2$	$(1,\pm i,0)$	$C_4^{(z)} \wedge \mathcal{T}_2^{(y)}$	$S_4^{(z)} \wedge \mathcal{T}_2^{(y)}$
		$(1,0,\pm i)$	$C_4^{(y)} \wedge \mathcal{T}_2^{(x)}$	$S_4^{(y)} \wedge \mathcal{T}_2^{(x)}$
		$(0,1,\pm i)$	$C_4^{(x)} \wedge \mathcal{T}_2^{(y)}$	$S_4^{(x)} \wedge \mathcal{T}_2^{(y)}$
$C_3 \wedge \mathcal{T}_s$	$C_3 \wedge \mathcal{T}_2^{(d)}$	$(1,\eta^2,\eta)$, $(1,\bar\eta,\bar\eta^2)$	$C_3^{(a)} \wedge \mathcal{T}_2^{\perp}$	$C_3^{(a)} \wedge \mathcal{T}_s^{\parallel}$
		$(1,\eta,\eta^2)$, $(1,\bar\eta^2,\bar\eta)$	$C_3^{(b)} \wedge \mathcal{T}_2^{\perp}$	$C_3^{(b)} \wedge \mathcal{T}_s^{\parallel}$
		$(1,\eta^2,\bar\eta^2)$, $(1,\bar\eta^2,\eta^2)$	$C_3^{(c)} \wedge \mathcal{T}_2^{\perp}$	$C_3^{(c)} \wedge \mathcal{T}_s^{\parallel}$
		$(1,\eta,\bar\eta)$, $(1,\bar\eta,\eta)$	$C_3^{(d)} \wedge \mathcal{T}_2^{\perp}$	$C_3^{(d)} \wedge \mathcal{T}_s^{\parallel}$
$C_{2v} \times \mathcal{T}$	$D_2' \times \mathcal{T}$	$(1,\pm 1,0)$	$D_2'^{(z)} \times \mathcal{T}$	$C_{2v}^{(z)} \times \mathcal{T}$
		$(1,0,\pm 1)$	$D_2'^{(y)} \times \mathcal{T}$	$C_{2v}^{(y)} \times \mathcal{T}$
		$(0,1,\pm 1)$	$D_2'^{(x)} \times \mathcal{T}$	$C_{2v}^{(x)} \times \mathcal{T}$

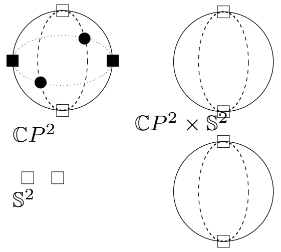
the case of the C_4, C_3, and D_2 orbits, one stabilizer corresponds to two different orbit points. Zero-dimensional strata of the action of the symmetry group $O \times \mathcal{T}$ and $T_d \times \mathcal{T}$ (where $\mathcal{T}$ acts as in (3.1a)) remain the same, but their stabilizers become larger. The order of the symmetry group $O \times \mathcal{T}$ is twice that of O and the order of all stabilizers is doubled. The structure of critical orbits also remains exactly the same except for the stabilizers. Each zero-dimensional stratum again consists of one orbit.

At the same time, two-dimensional invariant topological spheres $\mathbb{S}^2$ with stabilizers $C_2^{(a)}$, $a = \{x, y, z\}$, and $C_s^{(\alpha)}$, $\alpha = 1, \ldots, 6$, of the T_d action on $\mathbb{C}P^2$ (see Figure 10) become further stratified due to the action of the $\mathcal{T}$-extended group. Below we detail the action of $O \times \mathcal{T}$ (and $T_d \times \mathcal{T}$) on these invariant manifolds.

Stratification of each of the three C_2-invariant $\mathbb{S}^2$ spheres is equivalent to the natural action of the D_{2h} point symmetry group (see Figure 11, left). The generic two-dimensional

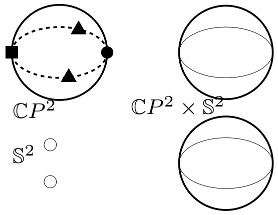
[10] We give notation for T_d, and O or O_h groups.

[11] Coordinates in terms of (z_1, z_2, z_3), $\eta = \exp(i\pi/3)$.



symbol	point stabilizer
□	$S_4^\alpha \wedge \mathcal{T}_s$ (C_{4v})
■	$D_{2d}^{\beta\neq\alpha} \wedge \mathcal{T}$ (D_4)
●	$C_{2v}^\alpha \wedge \mathcal{T}$ (D_{2h})
——	$C_2^\alpha \wedge \mathcal{T}_2^\beta$ (C_{2v})
- - - - -	$C_2^\alpha \wedge \mathcal{T}_s$ (C'_{2v})
· · · · ·	$C_2^\alpha \wedge \mathcal{T}$ (C_{2h})
surface	generic C_2^α

Figure 11. *Correlation between the C_2^a-invariant sphere on the $\mathbb{C}P^2$ space and C_2^a-invariant spheres on the $\mathbb{C}P^2 \times \mathbb{S}^2$ space whose coordinates on the $\mathbb{S}^2$ space (rotational sphere) are fixed to those of the $S_4^a \wedge \mathcal{T}_s$ points.*



symbol	point stabilizer
■	$D_{2d} \wedge \mathcal{T}$ (D_4)
▲	$C_{3v} \wedge \mathcal{T}$ (D_{3d})
●	$C_{2v} \wedge \mathcal{T}$ (D_{2h})
○	$C_s \wedge \mathcal{T}_2$ on $\mathbb{S}^2$
——	$C_s \wedge \mathcal{T}_2$ (C''_{2v})
- - - - -	$C_s \wedge \mathcal{T}$ (C'_{2h})
surface	generic C_s

Figure 12. *Correlation between the $C_s^{(k)}$-invariant sphere on the $\mathbb{C}P^2$ space and $C_s^{(k)}$-invariant spheres on the $\mathbb{C}P^2 \times \mathbb{S}^2$ space whose coordinates on the $\mathbb{S}^2$ space (rotational sphere) are fixed to those of the $C_s^{(k)} \wedge \mathcal{T}_2$ points.*

stratum has stabilizer C_2; three isolated (critical) two-point orbits with stabilizers $D_{2d}\wedge\mathcal{T}$, $S_4\wedge\mathcal{T}_2$, and $C_{2v}\wedge\mathcal{T}$ are described in Table 10; three one-dimensional strata with stabilizers $C_2\wedge g$, where $g = \mathcal{T}$, $\mathcal{T}_2$, or $\mathcal{T}_s$, are listed in Table 11. Together with isolated fixed points, these strata form $C_2\wedge g$-invariant circles $\mathbb{S}^1$.

Stratification of the six topological $\mathbb{S}^2$ spheres with stabilizers $C_s^{(\alpha)}$ deserves special comment. Each such sphere has one exceptional two-point critical orbit with stabilizer $C_{3v}\wedge\mathcal{T}$ (see Figure 12, left). This orbit cannot be found if we consider only those symmetry operations that act within the $C_s^{(\alpha)}$-invariant sphere, its presence is due to the action of the symmetry group $T_d \times \mathcal{T}$ on the entire $\mathbb{C}P^2$ space. Disregarding the exceptional $C_{3v}\wedge\mathcal{T}$ orbit, we can identify the action of the symmetry group $T_d \times \mathcal{T}$ within each $C_s^{(\alpha)}$ sphere with the natural action of the C_{2v} point group (on a sphere in a 3-space). This action has two critical one-point orbits with global stabilizers $D_{2d}\wedge\mathcal{T}$ and $C_{2v}\wedge\mathcal{T}$ (Table 10), which lie on the C_2 axis (horizontal axis of the leftmost sphere in Figure 12) and two different one-dimensional families of orbits with stabilizers $C_s\wedge g$, where $g = \mathcal{T}$ or $\mathcal{T}_2$ (Table 11). Except for the C_{2v} and D_{2d} points, all other points form two-point orbits of the spatial subgroup C_2. The $C_s\wedge g$ invariant circles combine respective one-dimensional strata and fixed points.

All two-dimensional invariant subspaces of the $T_d \times \mathcal{T}$ group action on $\mathbb{C}P^2$ are described in Table 12. In addition to the two types of invariant spheres $\mathbb{S}^2$, this action has a number of other two-dimensional invariant subspaces whose stabilizers $\mathcal{T}$, $\mathcal{T}_2$, and $\mathcal{T}_s$ include reversing symmetries. Generic stratum on the $\mathcal{T}$, $\mathcal{T}_2$, and $\mathcal{T}_s$ invariant subspaces includes orbits with

Table 11

One-dimensional strata on the $\mathbb{C}P^2$ *space under the same group action as in Table* 10.

Orbit stabilizer[12]		Coordinates[13]	Point stabilizer[12]	
T_d	O or O_h	on $\mathbb{C}P^2$	O or O_h	T_d
$C_2 \wedge \mathcal{T}$	$C_2 \wedge \mathcal{T}$	$(1, \pm a, 0)$	$C_2^{(z)} \wedge \mathcal{T}$	$C_2^{(z)} \wedge \mathcal{T}$
		$(1, \pm 1/a, 0)$		
		$(1, 0, \pm a)$	$C_2^{(y)} \wedge \mathcal{T}$	$C_2^{(y)} \wedge \mathcal{T}$
		$(1, 0, \pm 1/a)$		
		$(0, 1, \pm a)$	$C_2^{(x)} \wedge \mathcal{T}$	$C_2^{(x)} \wedge \mathcal{T}$
		$(0, 1, \pm 1/a)$		
$C_2 \wedge \mathcal{T}_2$	$C_2 \wedge \mathcal{T}_2$	$(1, \pm ia, 0)$	$C_2^{(z)} \wedge \mathcal{T}_2^{(x)}$	$C_2^{(z)} \wedge \mathcal{T}_2^{(x)}$
		$(1, \pm i/a, 0)$		
		$(1, 0, \pm ia)$	$C_2^{(y)} \wedge \mathcal{T}_2^{(z)}$	$C_2^{(y)} \wedge \mathcal{T}_2^{(z)}$
		$(1, 0, \pm i/a)$		
		$(0, 1, \pm ia)$	$C_2^{(x)} \wedge \mathcal{T}_2^{(y)}$	$C_2^{(x)} \wedge \mathcal{T}_2^{(y)}$
		$(0, 1, \pm i/a)$		
$C_2 \wedge \mathcal{T}_s$	$C_2 \wedge \mathcal{T}_2^{(d)}$	$(1, \pm \eta, 0)$	$C_2^{(z)} \wedge \mathcal{T}_2^{(d)}$	$C_2^{(z)} \wedge \mathcal{T}_s^{\parallel}$
		$(1, \pm \bar{\eta}, 0)$		
		$(1, 0, \pm \eta)$	$C_2^{(y)} \wedge \mathcal{T}_2^{(d)}$	$C_2^{(y)} \wedge \mathcal{T}_s^{\parallel}$
		$(1, 0, \pm \bar{\eta})$		
		$(0, 1, \pm \eta)$	$C_2^{(x)} \wedge \mathcal{T}_2^{(d)}$	$C_2^{(x)} \wedge \mathcal{T}_s^{\parallel}$
		$(0, 1, \pm \bar{\eta})$		
$C_s \wedge \mathcal{T}$	$C_2^{(d)} \wedge \mathcal{T}$	$(1, 1, \pm a)$	$C_2^{(1)} \wedge \mathcal{T}$	$C_s^{(xy)} \wedge \mathcal{T}$
		$(1, -1, \pm a)$	$C_2^{(2)} \wedge \mathcal{T}$	$C_s^{(\overline{xy})} \wedge \mathcal{T}$
		$(1, \pm a, 1)$	$C_2^{(3)} \wedge \mathcal{T}$	$C_s^{(xz)} \wedge \mathcal{T}$
		$(1, \pm a, -1)$	$C_2^{(4)} \wedge \mathcal{T}$	$C_s^{(\overline{xz})} \wedge \mathcal{T}$
		$(1, \pm a, \pm a)$	$C_2^{(5)} \wedge \mathcal{T}$	$C_s^{(yz)} \wedge \mathcal{T}$
		$(1, \pm a, \mp a)$	$C_2^{(6)} \wedge \mathcal{T}$	$C_s^{(\overline{yz})} \wedge \mathcal{T}$
$C_s \wedge \mathcal{T}_2$	$C_2^d \wedge \mathcal{T}_2$	$(1, 1, \pm ia)$	$C_2^{(1)} \wedge \mathcal{T}_2^{(z)}$	$C_s^{(xy)} \wedge \mathcal{T}_2^{(z)}$
		$(1, -1, \pm ia)$	$C_2^{(2)} \wedge \mathcal{T}_2^{(z)}$	$C_s^{(\overline{xy})} \wedge \mathcal{T}_2^{(z)}$
		$(1, \pm ia, 1)$	$C_2^{(3)} \wedge \mathcal{T}_2^{(y)}$	$C_s^{(xz)} \wedge \mathcal{T}_2^{(y)}$
		$(1, \pm ia, -1)$	$C_2^{(4)} \wedge \mathcal{T}_2^{(y)}$	$C_s^{(\overline{xz})} \wedge \mathcal{T}_2^{(y)}$
		$(1, \pm ia, \pm ia)$	$C_2^{(5)} \wedge \mathcal{T}_2^{(x)}$	$C_s^{(yz)} \wedge \mathcal{T}_2^{(x)}$
		$(1, \pm ia, \mp ia)$	$C_2^{(6)} \wedge \mathcal{T}_2^{(x)}$	$C_s^{(\overline{yz})} \wedge \mathcal{T}_2^{(x)}$

24, 8, and 4 points, respectively. The topology of all these subspaces is $\mathbb{R}P^2$.

4.4. Rotational structure of the E mode: Action on $\mathbb{C}P^1 \times \mathbb{S}^2 \sim \mathbb{S}^2 \times \mathbb{S}^2$. The Hamiltonian that describes rotational structure of the E-mode polyads is defined on the four-dimensional reduced rotation–vibration phase space $\mathbb{C}P^1 \times \mathbb{S}^2$, the direct product of the vibrational E-mode polyad space $\mathbb{C}P^1$ (polyad sphere) and rotational sphere $\mathbb{S}^2$. We use

[12] We give notation for either the T_d, O, or O_h group.

[13] Here $a \neq 0, \pm 1, \infty$, and $\eta = e^{i\varphi}$, where $\varphi \neq k\pi/2$.

Table 12

Two-dimensional invariant subspaces of $\mathbb{C}P^2$ under the same group action as in Tables 10 and 11.

Orbit stabilizer[14]		Coordinates[15]	Point stabilizer[14]		Topology
T_d	O or O_h	on $\mathbb{C}P^2$	O or O_h	T_d	
C_2	C_2	$(1, w, 0)$	$C_2^{(z)}$	$C_2^{(z)}$	$\mathbb{S}^2$
		$(1, 0, w)$	$C_2^{(y)}$	$C_2^{(y)}$	
		$(0, 1, w)$	$C_2^{(x)}$	$C_2^{(x)}$	
C_s	$C_2^{(d)}$	$(1, 1, w)$	$C_2^{(1)}$	$C_s^{(xy)}$	$\mathbb{S}^2$
		$(1, -1, w)$	$C_2^{(2)}$	$C_s^{(\overline{xy})}$	
		$(1, w, 1)$	$C_2^{(3)}$	$C_s^{(xz)}$	
		$(1, w, -1)$	$C_2^{(4)}$	$C_s^{(\overline{xz})}$	
		$(1, w, w)$	$C_2^{(5)}$	$C_s^{(yz)}$	
		$(1, w, -w)$	$C_2^{(6)}$	$C_s^{(\overline{yz})}$	
$\mathcal{T}$	$\mathcal{T}$	$(1, a, b)$	$\mathcal{T}$	$\mathcal{T}$	$\mathbb{R}P^2$
$\mathcal{T}_2$	$\mathcal{T}_2$	$(1, a, ib)$	$\mathcal{T}_2^{(z)}$	$\mathcal{T}_2^{(z)}$	$\mathbb{R}P^2$
		$(1, ia, b)$	$\mathcal{T}_2^{(y)}$	$\mathcal{T}_2^{(y)}$	
		$(1, ia, ib)$	$\mathcal{T}_2^{(x)}$	$\mathcal{T}_2^{(x)}$	
$\mathcal{T}_s$	$\mathcal{T}_2^{(d)}$	$(1, w, w)$	$\mathcal{T}_2^{(d_1)}$	$\mathcal{T}_s^{(yz)}$	$\mathbb{R}P^2$
		$(1, w, -w)$	$\mathcal{T}_2^{(d_2)}$	$\mathcal{T}_s^{(\overline{yz})}$	
		$(1, a\eta, \eta^2)$	$\mathcal{T}_2^{(d_3)}$	$\mathcal{T}_s^{(xz)}$	
		$(1, a\eta, -\eta^2)$	$\mathcal{T}_2^{(d_4)}$	$\mathcal{T}_s^{(\overline{xz})}$	
		$(1, \eta^2, a\eta)$	$\mathcal{T}_2^{(d_5)}$	$\mathcal{T}_s^{(xy)}$	
		$(1, -\eta^2, a\eta)$	$\mathcal{T}_2^{(d_6)}$	$\mathcal{T}_s^{(\overline{xy})}$	

information on the stratification of the individual factor spaces $\mathbb{C}P^1$ (section 4.2) and $\mathbb{S}^2$ (section 4.1) in order to find the stratification of $\mathbb{C}P^1 \times \mathbb{S}^2$.

Let (v) and (r) be points on $\mathbb{C}P^1$ and $\mathbb{S}^2$, respectively, and let (v, r) denote points on the rovibrational (i.e., rotational-vibrational) space $\mathbb{C}P^1 \times \mathbb{S}^2$. The stabilizer $G_{v,r}$ of point (v, r) is an *intersection* $G_v \cap G_r$ of stabilizers on $\mathbb{C}P^1$ and $\mathbb{S}^2$. In simple terms, the symmetry of (v, r) can only be *lower* than that of its projections (v) and (r). The dimension of the stratum $\{v, r\}$ on $\mathbb{C}P^1 \times \mathbb{S}^2$ is the sum of the dimensions of strata $\{v\}$ and $\{r\}$. The stratum $\{v, r\}$ on the product space is connected if both its projections $\{v\}$ and $\{r\}$ on the two factor subspaces are connected.

Most important and basic to our analysis are critical orbits (v, r). Such orbits can either be nonconnected parts of a nonzero-dimensional stratum or belong to a stratum $\{v, r\}$ of dimension zero. These latter strata occur when both strata $\{r\}$ and $\{v\}$ have dimension zero *and* there is no stratum of nonzero dimension with stabilizer $G_r \cap G_v$.

Using the lattice of conjugate subgroups of $T_d \times \mathcal{T}$ in Figure 8, where we indicated all possible stabilizers G_v and G_r, the reader can easily find the intersections $G_v \cap G_r$ and then look up the details of the structure of the particular subgroups in Table 2. Indeed, all subgroups of a given stabilizer G are found by descending along the lattice paths which originated

[14] We give notation for either the T_d, O, or O_h group.

[15] Here a, b are real, w is a complex number, and $\eta = e^{i\varphi}$.

Table 13

Intersection (correlation) of stabilizers of the $T_d \times \mathcal{T}$ group action on the E-mode polyad phase space $\mathbb{C}P^1$ (left column) and on the rotational phase space $\mathbb{S}^2$ (top row). For information on the notation and structure of subgroups of $T_d \times \mathcal{T}$ refer to Table 2. Critical orbits on $\mathbb{C}P^1 \times \mathbb{S}^2$ are underlined.

			Strata[16] on the rotational phase space $\mathbb{S}^2$					
			0D	0D	0D	1D	1D	2D
Strata[16] on			(4)	(3)	(6)	(3)	(6)	
the $\mathbb{C}P^1$ space			$C_3 \wedge \mathcal{T}_s$	$S_4 \wedge \mathcal{T}_s$	$C_s \wedge \mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_s$	C_1
0D		$T \wedge \mathcal{T}_s$	$\underline{C_3^{(k)} \wedge \mathcal{T}_s}$	$C_2^{\xi} \wedge \mathcal{T}_s$	$\mathcal{T}_s^{\xi_2}$	C_1	$\mathcal{T}_s^{\beta}$	C_1
0D	(3)	$D_{2d} \times \mathcal{T}$	$\mathcal{T}_s$	$\underline{S_4^{\eta} \wedge \mathcal{T}_s}$[17] $C_2^{\eta} \wedge \mathcal{T}_2^{\xi}$[20]	$\underline{C_s \wedge \mathcal{T}_2}$[19] $\mathcal{T}_2^{\xi}$[20]	$\mathcal{T}_2^{\eta}$	$\mathcal{T}_s$[19] C_1[20]	C_1
1D	(3)	$D_2 \wedge \mathcal{T}_s$	$\mathcal{T}_s$	$C_2^{\eta} \wedge \mathcal{T}_s$[17] C_2^{ξ}[18]	$\mathcal{T}_s$[19] C_1[20]	C_1	$\mathcal{T}_s$[19] C_1[20]	C_1
1D		$D_2 \times \mathcal{T}$	C_1	$C_2^{\eta} \wedge \mathcal{T}_2^{\xi}$[20]	$\mathcal{T}_2$	$\mathcal{T}_2^{\xi}$	C_1	C_1
2D		D_2	C_1	C_2^{ξ}	C_1	C_1	C_1	C_1

at G. The highest node, where the paths descending from G_v and G_r join, is the intersection $G_v \cap G_r$, i.e., the largest common subgroup of G_v and G_r. All possible intersections are summarized in Table 13. We comment on them below starting with the low-symmetry strata.

4.4.1. Noncritical orbits. Generic points (r) on $\mathbb{S}^2$ have stabilizer $G_r = C_1$ (rightmost column in Table 13). Obviously, these points lift to points (v, r) with trivial symmetry C_1 regardless of the stabilizer G_v. The generic stratum $\{v, r\}$ is of dimension four, has stabilizer C_1, and includes 48-point orbits.

Generic points (v) on $\mathbb{C}P^1$ (bottom row in Table 13) have stabilizer D_2, which has three order-two subgroups C_2^{η} with axes $\eta = \{x, y, z\}$. Intersection of D_2 with stabilizers G_r equals C_1 in all cases except for the three conjugate subgroups $G_r^{\eta} = S_4^{\eta} \wedge \mathcal{T}_s$. Intersection $G_r^{\eta} \cap D_2$ is the particular C_2^{η} subgroup of D_2. The corresponding C_2 stratum on $\mathbb{C}P^1 \times \mathbb{S}^2$ includes 24-point orbits. Since $S_4^{\eta} \wedge \mathcal{T}_s$ points on $\mathbb{S}^2$ are critical (fixed), this stratum has dimension two.

The stabilizer $G_v = D_2 \wedge \mathcal{T}$ of the one-dimensional stratum on $\mathbb{C}P^1$ is an invariant subgroup of $T_d \times \mathcal{T}$; its intersection with subgroups G_r of the same class of conjugate subgroups of $T_d \times \mathcal{T}$ results again in conjugate subgroups. For $G_r = S_4 \wedge \mathcal{T}_s$ the intersection is $C_2 \times \mathcal{T}_2$ and the stratum $\{v, r\}$ has dimension one. If $G_r = \mathcal{T}_2^b$, where $b = \{x, y, z\}$, we have a two-dimensional

[16] For each space, strata of dimensions zero, one, and two are marked as 0D, 1D, and 2D; the number of different conjugate stabilizers in the same class of $T_d \times \mathcal{T}$ is given in parentheses. Right superscripts define concrete conjugate stabilizers: η and ξ label axes (x, y, z) for orbits on $\mathbb{C}P^1$ and $\mathbb{S}^2$, respectively; (k) and β distinguish stabilizers of orbits on $\mathbb{S}^2$ in the C_3 class and the σ class; they can take four and six values, respectively.

[17] $\eta = \xi$; axes η and ξ are the *same*.

[18] $\eta \neq \xi$; axes η and ξ are *different*.

[19] $\beta = \{\eta_1, \eta_2\}$; stabilizer of the orbits on $\mathbb{S}^2$ of index (β) includes one of the two operations of index $\{\eta_1, \eta_2\}$; see text for examples.

[20] $\beta \neq \{\eta_1, \eta_2\}$; stabilizer of the orbits on $\mathbb{S}^2$ of index (β) does not include operations of index $\{\eta_1, \eta_2\}$.

family $\{v, r\}$ of 24-point orbits with symmetry $\mathcal{T}_2^b$. If $G_r = C_s \wedge \mathcal{T}_2$, the intersection is again $\mathcal{T}_2$; i.e., isolated fixed points (r) on $\mathbb{S}^2$ with stabilizer $C_s \wedge \mathcal{T}_2$ lift to points (v, r) with stabilizer $\mathcal{T}_2$. Since the zero-dimensional stratum $C_s \wedge \mathcal{T}_2$ on $\mathbb{S}^2$ is the closure of the one-dimensional stratum $\mathcal{T}_2$ on $\mathbb{S}^2$, the entire $\mathcal{T}_2$ stratum $\{v, r\}$ is connected. As a consequence, isolated points (r) with stabilizer $C_s \wedge \mathcal{T}_2$ lift to noncritical points (v, r) on the two-dimensional stratum on $\mathbb{C}P^1 \times \mathbb{S}^2$.

When both G_v and G_r are not invariant subgroups, the symmetry of the $\{v, r\}$ stratum depends on the choice of the subgroup within the class of conjugate subgroups. In the case of $G_v = D_2 \wedge \mathcal{T}_s$ and $G_r = \mathcal{T}_s$ the intersection can be either $\mathcal{T}_s$ or C_1. As shown in Table 2, the element C_2^a, where $a = \{x, y, z\}$, of $D_2 \wedge \mathcal{T}_s$ distinguishes a particular stabilizer in the class of three conjugate subgroups. This stabilizer has two $\mathcal{T}_s$ subgroups generated by $\mathcal{T}_s^{a_1}$ and $\mathcal{T}_s^{a_2}$. If, therefore, G_r is one of these two subgroups, then its intersection with G_v is nontrivial; otherwise $G_v \cap G_r = C_1$. In the case of $G_v = D_2 \wedge \mathcal{T}_s$ and $G_r = \mathcal{T}_2$ the intersection is always trivial. On the other hand, intersection of $G_v = D_2 \wedge \mathcal{T}_s$ and $C_3 \wedge \mathcal{T}_s$ is always a nontrivial subgroup $\mathcal{T}_s$ because one of the two orthogonal symmetry planes $\mathcal{T}_s^{a_{1,2}}$ in $D_2 \wedge \mathcal{T}_s$ always contains the particular axis C_3 of $C_3 \wedge \mathcal{T}_s$. In other words, intersection of the set of two $\mathcal{T}_s$ elements in $D_2 \wedge \mathcal{T}_s$ and three $\mathcal{T}_s$ elements in $C_3 \wedge \mathcal{T}_s$ (which cross on the C_3 axis) is never empty. Observe that the $\mathcal{T}_s$ stratum $\{v, r\}$ has dimension two and is connected. All points in Table 13 that lift to this stratum become noncritical; the $\{v, r\}$ parts of dimensions zero and one form the closure.

4.4.2. Critical orbits (strata of dimension zero).

Orbits that project on zero-dimensional strata of $\mathbb{C}P^1$ and $\mathbb{S}^2$ *and* have unique stabilizers are critical. The six critical orbits on $\mathbb{C}P^1 \times \mathbb{S}^2$ are characterized below.

Stabilizer on $\mathbb{C}P^1$	$T \wedge \mathcal{T}_s$	$D_{2d} \times \mathcal{T}$	$D_{2d} \times \mathcal{T}$
Stabilizer $\mathbb{C}P^1 \times \mathbb{S}^2$	$C_3 \wedge \mathcal{T}_s$	$S_4 \wedge \mathcal{T}_s$	$C_s \times \mathcal{T}_2$
Number of orbits	2	2	2
Number of points in orbit	8	6	12

We explain how to find these orbits using Table 13.

Consider the stratum $\{v\}$ with stabilizer $G_v = T \wedge \mathcal{T}_s$, which consists of one two-point orbit, and the stratum $\{r\}$ with stabilizer $G_r = C_3 \wedge \mathcal{T}_s$, which consists of one eight-point orbit. Since each of the four conjugate subgroups G_r is a subgroup of G_v, we have the zero-dimensional stratum $\{v, r\}$ with stabilizer $C_3 \wedge \mathcal{T}_s$, which includes two eight-point orbits (all points on the orbit $\{r\}$ lift to the same orbit $\{v, r\}$). If for the same $\{v\}$ we consider $\{r\}$ with stabilizer $G_r = S_4 \wedge \mathcal{T}_s$, the resulting 12-point orbit $\{v, r\}$ has the stabilizer $C_2 \times \mathcal{T}_s$ and should, therefore, be part of a one-dimensional stratum. Indeed, as can be seen from Figure 8,

$$T \wedge \mathcal{T}_s \cap S_4 \wedge \mathcal{T}_s = D_2 \wedge \mathcal{T}_s \cap S_4 \wedge \mathcal{T}_s = C_2 \times \mathcal{T}_s.$$

Consider now the stratum $\{v\}$ with stabilizer $G_v = D_{2d} \times \mathcal{T}$, which again consists of one two-point orbit. The intersection of G_v with $G_r = S_4 \wedge \mathcal{T}_s$ can be the whole G_r (i.e., G_r can be the subgroup of G_v) if both subgroups include the same element C_2^η, where axis η can be x, y, or z. In this case we have a zero-dimensional stratum $\{v, r\}$ with stabilizer $S_4 \wedge \mathcal{T}_s$, which includes two six-point orbits (all points on the six-point orbit $\{r\}$ lift to the same orbit $\{v, r\}$). If for the same $\{v\}$ we take $\{r\}$ with stabilizer $G_r = C_s \times \mathcal{T}_2$, this latter stabilizer can

Table 14

Intersection (correlation) of stabilizers of the T_d group action on the F_2-mode vibrational phase space $\mathbb{C}P^2$ (left column), on the rotational phase space $\mathbb{S}^2$ (top row, center), and on the E-mode vibrational phase space $\mathbb{C}P^1$ (top row, right). Critical orbits are underlined.

		Strata[21] on $\mathbb{S}^2$				Strata[21] on $\mathbb{C}P^1$		
Strata[21] on $\mathbb{C}P^2$		0D	0D	0D	2D	0D	0D	2D
		$S_4^{(\eta')}$	$C_3^{(k')}$	$C_s^{(\beta')}$	C_1	T	$D_{2d}^{(\eta')}$	D_2
		$\eta'\leq3$	$k'\leq4$	$\beta'\leq6$			$\eta'\leq3$	
0D	$D_{2d}^{(\eta)}$ $\eta\leq3$	$\underline{S_4}$[22] C_2[23]	C_1	C_s[24] C_1	C_1	D_2	$\underline{D_{2d}}$[22] D_2[23]	D_2
0D	$S_4^{(\eta)}$ $\eta\leq3$	$\underline{S_4}$[22] C_1[23]	C_1	C_1	C_1	C_2	$\underline{S_4}$[22] C_2[23]	C_2
0D	$C_{3v}^{(k)}$ $k\leq4$	C_1	$\underline{C_3}$[22] C_1[23]	C_s[24] C_1	C_1	$\underline{C_3}$	C_s	C_1
0D	$C_3^{(k)}$ $k\leq4$	C_1	$\underline{C_3}$[22] C_1[23]	C_1	C_1	$\underline{C_3}$	C_1	C_1
0D	$C_{2v}^{(\eta)}$ $\eta\leq3$	C_2[22] C_1[23]	C_1	$\underline{C_s}$[24,25] C_s[24] C_1	C_1	C_2	$\underline{C_{2v}}$[22] C_2[23]	C_2
2D	$C_2^{(\eta)}$ $\eta\leq3$	C_2[22] C_1[23]	C_1	C_1	C_1	C_2	C_2[22] C_2[23]	C_2
2D	$C_s^{(\beta)}$ $\beta\leq6$	C_1	C_1	C_s[22] C_1[23]	C_1	C_1	C_s[24] C_1	C_1
4D	C_1	C_1	C_1	C_1	C_1	C_1	C_1	C_1

again be a subgroup of G_v if both share the same $C_s = \sigma$ element, which can be either σ^{η_1} or σ^{η_2} (see Table 2). Since in this case $\{r\}$ consists of one 12-point orbit, the corresponding zero-dimensional $\{v, r\}$ stratum with stabilizer $C_s \times \mathcal{T}_2$ contains two 12-point orbits.

4.5. Rotational structure of the F_2 mode: Action of T_d and $T_d \times \mathcal{T}$ on $\mathbb{C}P^2 \times \mathbb{S}^2$.

We now combine small-amplitude F_2-mode vibrations and rotation. The Hamiltonian of this system is a $(T_d \times \mathcal{T})$-invariant function on the six-dimensional reduced phase space $\mathbb{C}P^2 \times \mathbb{S}^2$. The action of the full symmetry group $T_d \times \mathcal{T}$ on $\mathbb{C}P^2 \times \mathbb{S}^2$ can be found from that on the individual spaces $\mathbb{C}P^2$ and $\mathbb{S}^2$ using the approach of the previous section; in particular, we

[21] Strata of dimension s are marked as sD, $s = 0, 1, 2, 4$. Classes of stabilizers of the strata are listed in T_d notation. Indexes η, k, β distinguish different stabilizers on $\mathbb{C}P^2$ within the same class; indexes η', k', β' are used for $\mathbb{S}^2$ or $\mathbb{C}P^1$. Refer to Table 2 for information on the notation and structure of subgroups of $T_d \times \mathcal{T}$.

[22] Identical subgroups of $T_d \times \mathcal{T}$: axes η and η', k and k', or subgroups β and β' are the *same*, i.e., $\eta \equiv \eta'$, $k \equiv k'$, or $\beta \equiv \beta'$.

[23] Different subgroups of $T_d \times \mathcal{T}$ of the same class: axes η and η', k and k', or subgroups β and β' are *different*, i.e., $\eta \neq \eta'$, etc.

[24] Subgroups $D_{2d}^{(\eta)}$, $C_{2v}^{(\eta)}$, $C_{3v}^{(k)}$ include the symmetry plane β'.

[25] β' equals η_1 or η_2; the point projects on the disconnected zero-dimensional component of the $C_s^{\eta_1}$ or $C_s^{\eta_2}$ stratum on $\mathbb{C}P^2$; see section 4.5.1.

determine intersections of stabilizers from the subgroup lattice of $T_d \times \mathcal{T}$.

Essential information can be obtained from the simpler study of the action of the spatial symmetry group T_d whose subgroup lattice is given in Figure 1. As shown in Table 14, the action of T_d on the rotation–vibration space $\mathbb{C}P^2 \times \mathbb{S}^2$ creates strata of symmetry S_4, C_2, C_3, C_s, and C_1 (generic). Some strata, notably C_s, have disconnected components of different dimension. By dimension of these strata, we mean the highest dimension of their components. The C_1 stratum has dimension six, components of other strata can be of dimension zero and two.

4.5.1. Strata and components of dimension zero (critical orbits).

The S_4 stratum on $\mathbb{C}P^2 \times \mathbb{S}^2$ projects on the S_4 orbits on the rotational space $\mathbb{S}^2$ and on the D_{2d} or S_4 orbits on the vibrational space $\mathbb{C}P^2$. Each $G_r = S_4^{(\eta')}$ stabilizer has two fixed points on $\mathbb{S}^2$ (which belong to the same orbit of the T_d action). Each $G_v = D_{2d}^{(\eta)}$ stabilizer has one fixed point on $\mathbb{C}P^2$ (see Table 10). Consequently, there are two points on $\mathbb{C}P^2 \times \mathbb{S}^2$ with the same stabilizer

$$G_r \cap G_v = D_{2d}^{(\eta)} \cap S_4^{(\eta)} = S_4^{(\eta)}.$$

The six points corresponding to three different conjugate stabilizers with $\eta = x, y, z$ form one six-point orbit, which we label $A^{(4)}$. If for the same G_r we take $G_v = S_4^{(\eta)}$, we combine two points on $\mathbb{C}P^2$ (which are in the same orbit) with two points on $\mathbb{S}^2$. It is important to observe that the resulting four points on $\mathbb{C}P^2 \times \mathbb{S}^2$ belong to two different orbits: one pair belongs to orbit $B^{(4)}$ and another to orbit $C^{(4)}$. With three possible axes η taken into account, orbits B and C contain six points each.

The above description of the S_4 orbits $B^{(4)}$ and $C^{(4)}$ can be easily verified on a concrete example. Such examples are given in Table 15, where the E-mode coordinates (z_4, z_5) should at present be ignored. Consider a particular axis $\eta = z$ (axis 3). The two fixed points on the $\mathbb{C}P^2$ space have coordinates (Table 10)

$$(z_1, z_2, z_3) = (1, \pm i, 0), \tag{4.1a}$$

and the coordinates of the two fixed points on the $\mathbb{S}^2$ sphere are

$$(j_1, j_2, j_3) = (0, 0, \pm 1). \tag{4.1b}$$

(For simplicity, here and in the rest of this section, we drop normalization factors, which are not essential to our current discussion.) The group T_d acts on $\mathbb{C}P^2$ and $\mathbb{S}^2$ in such a way that any operation in T_d that interchanges the two points (4.1a) on $\mathbb{C}P^2$ interchanges the two points (4.1b) on $\mathbb{S}^2$. The C_2 rotation about axis x (axis 1) is an example:

$$C_2^x : (z_1, z_2, z_3; j_1, j_2, j_3) \to (z_1, -z_2, -z_3; j_1, -j_2, -j_3).$$

As a result, no operation in T_d maps points

$$B = (1, \pm i, 0; 0, 0, \pm 1) \quad \text{and} \quad C = (1, \pm i, 0; 0, 0, \mp 1)$$

of the $\mathbb{C}P^2 \times \mathbb{S}^2$ space into each other; these points, therefore, belong to *different* orbits.

Table 15

Fixed points of the $T_d \times \mathcal{T}$ group action on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$.

(a) Stabilizer S_4^z (or C_4), $r_f = \sqrt{N_f}$, $r_e = \sqrt{2N_e}$, $r_j = J$.

Point	$\frac{z_1}{r_f}$	$\frac{z_2}{r_f}$	$\frac{z_3}{r_f}$	$\frac{z_4}{r_e}$	$\frac{z_5}{r_e}$	$\frac{j_1}{r_j}$	$\frac{j_2}{r_j}$	$\frac{j_3}{r_j}$
A_1	0	0	$\sqrt{2}$	1	0	0	0	1
A_1'	0	0	$\sqrt{2}$	1	0	0	0	-1
A_2	0	0	$\sqrt{2}$	0	1	0	0	1
A_2'	0	0	$\sqrt{2}$	0	1	0	0	-1
B_1	1	i	0	1	0	0	0	1
B_1'	1	$-i$	0	1	0	0	0	-1
B_2	1	i	0	0	1	0	0	1
B_2'	1	$-i$	0	0	1	0	0	-1
C_1	1	i	0	1	0	0	0	-1
C_1'	1	$-i$	0	1	0	0	0	1
C_2	1	i	0	0	1	0	0	-1
C_2'	1	$-i$	0	0	1	0	0	1

(b) Stabilizer C_3 [111], $r_f = (\sqrt{2N_f})/\sqrt{3}$, $r_e = \sqrt{N_e}$, $r_j = \sqrt{3}J$, $\chi = \exp(2\pi i/3)$.

Point	$\frac{z_1}{r_f}$	$\frac{z_2}{r_f}$	$\frac{z_3}{r_f}$	$\frac{z_4}{r_e}$	$\frac{z_5}{r_e}$	$\frac{j_1}{r_j}$	$\frac{j_2}{r_j}$	$\frac{j_3}{r_j}$
A_1	1	1	1	1	i	1	1	1
A_2	1	1	1	1	i	-1	-1	-1
B_1	1	χ	$\bar{\chi}$	1	i	1	1	1
B_2	1	χ	$\bar{\chi}$	1	i	-1	-1	-1
C_1	1	χ	$\bar{\chi}$	1	i	-1	-1	-1
C_2	1	χ	$\bar{\chi}$	1	i	1	1	1

(c) Stabilizer C_s^{xy} (or C_2), $r_f = \sqrt{N_f}$, $r_e = \sqrt{2N_e}$, $r_j = J/\sqrt{2}$.

Point	$\frac{z_1}{r_f}$	$\frac{z_2}{r_f}$	$\frac{z_3}{r_f}$	$\frac{z_4}{r_e}$	$\frac{z_5}{r_e}$	$\frac{j_1}{r_j}$	$\frac{j_2}{r_j}$	$\frac{j_3}{r_j}$
A_1	1	-1	0	1	0	1	-1	0
A_2	1	-1	0	0	1	-1	1	0

The three eight-point critical orbits of the C_3 stratum can be described analogously; cf. Table 15(b). Combination of the $C_{3v}^{(k)}$ point on $\mathbb{C}P^2$ and two $C_3^{(k)}$ points on $\mathbb{S}^2$ gives two points on the $A^{(3)}$ orbit. With all four axes C_3 taken into account $(k = 1, \dots, 4)$, the orbit has eight points. The $B^{(3)}$- and $C^{(3)}$-type orbits of the C_3 action include points (z, j) and $(z, -j)$, respectively, where (z) and (j) stand for the coordinates of the fixed point on $\mathbb{C}P^2$ and $\mathbb{S}^2$.

The two critical fixed points of the $C_s^{(\alpha)}$ action on $\mathbb{C}P^2 \times \mathbb{S}^2$ are obtained when we combine the only isolated fixed point of the $C_s^{(\alpha)}$ action on $\mathbb{C}P^2$ (see Table 10) with the two respective points on $\mathbb{S}^2$ (recall that C_s acts both on $\mathbb{C}P^2$ and on $\mathbb{S}^2$ as operation C_2 of the O group). In the particular case of $C_{2v}^{(z)}$ and its subgroup $C_s^{(xy)}$ (set $\eta = z$ and $\beta' = \eta_1 = xy$ in Table 14), we combine the $z = (1, -1, 0)$ fixed point of the $C_s^{(xy)}$ action on $\mathbb{C}P^2$ with two points $j = \pm(1, -1, 0)$ on $\mathbb{S}^2$. For six conjugate elements C_s, we obtain a 12-point orbit, which is isolated from the rest of the C_s stratum (because its projection on the $\mathbb{C}P^2$ space is isolated). This orbit is therefore critical.

4.5.2. Strata of dimension two. Action of T_d on $\mathbb{C}P^2 \times \mathbb{S}^2$ creates two strata of dimension two, C_2 and C_s. The C_s stratum has a disconnected component of dimension zero, described above. In both cases, we combine isolated fixed points on the rotational space $\mathbb{S}^2$ and points on the invariant spheres $\mathbb{S}^2 \subset \mathbb{C}P^2$. This is illustrated in Figure 11 and 12.

We describe first the C_2 stratum. Since $C_2 = S_4^2$, points with local symmetry C_2 and S_4 coincide on the rotational space $\mathbb{S}^2_j$. In the particular example of $C_2^z = (S_4^z)^2$ these points are given in (4.1b). We combine points on $\mathbb{S}^2_j$ with points on the C_2-invariant sphere in $\mathbb{C}P^2$. In our C_2^z example, these latter points are (see Table 12)

$$(z_1, z_2, z_3) = (1, w, 0), \quad \operatorname{Re} w \neq 0.$$

All points on $\mathbb{C}P^2 \times \mathbb{S}^2$ with stabilizer C_2^z are

$$(z_1, z_2, z_3;\ j_1, j_2, j_3) = (1, w, 0;\ 0, 0, \pm 1), \quad \operatorname{Re} w \neq 0.$$

Removing the above restriction on w adds four critical points $(1, \pm i, 0;\ 0, 0, \pm 1)$ with stabilizer S_4^z and produces two spheres shown in Figure 11, right. These are the only isolated critical points which remain on the C_2-invariant spheres when we add rotation; all other fixed points which lie on the C_2-invariant sphere of the purely vibrational system with phase space $\mathbb{C}P^2$ (i.e., when $j = 0$) disappear. For example, consider the two D_{2d} points $z = (1, 0, 0)$ and $(0, 1, 0)$ with stabilizers D_{2d}^x and D_{2d}^y, respectively (Table 10 and Figure 11, left). Their rotational coordinates should necessarily be $j = (1, 0, 0)$ and $(0, 1, 0)$, which project to points on $\mathbb{S}^2$ with stabilizers S_4^x and S_4^y, respectively.

A similar approach can be used to describe the C_s-invariant spheres (the two-dimensional component of the C_s stratum) in the $\mathbb{C}P^2 \times \mathbb{S}^2$ space shown in Figure 12, right. In this case, *no* critical points remain on the spheres when rotation is added and we lift from $\mathbb{C}P^2$ to $\mathbb{C}P^2 \times \mathbb{S}^2$. As an example, consider the reflection plane $x = y$ whose action is given in Table 4. This operation has two fixed points on the rotational sphere $\mathbb{S}^2_j$:

$$(j_1, j_2, j_3) = (\pm 1, \mp 1, 0).$$

(At these points, axial vector j is orthogonal to the plane $x = y$.) Operation C_s^{xy} is a combination of inversion and rotation by π about axis $x = -y$, i.e., $(1, -1, 0)$. The points on the $\mathbb{C}P^1 \sim \mathbb{S}^2$ subspace orthogonal to this axis have, therefore, coordinates (see Table 12)

$$(z_1, z_2, z_3) = (1, 1, w),$$

and the points of the two C_s-invariant spheres $\mathbb{S}^2$ in the six-dimensional space $\mathbb{C}P^2 \times \mathbb{S}^2$ are

$$(z_1, z_2, z_3;\ j_1, j_2, j_3) = (1, 1, w;\ \pm 1, \mp 1, 0).$$

Stabilizers of critical points, which are present on the C_s-invariant sphere when $j = 0$ (Figure 12, left), are such that these points become noncritical when rotation is added. Thus, the C_{2v} point with $z = (1, 1, 0)$, which lies on the C_s^{xy}-invariant sphere of our example, has the stabilizer C_2^z and must, therefore, combine with $j = (0, 0, 1)$ in order to have a higher stabilizer.

4.5.3. Extension by time reversal $\mathcal{T}$. The action of the full group $T_d \times \mathcal{T}$ on the $\mathbb{C}P^2 \times \mathbb{S}^2$ space creates more strata. For our purposes, however, it is sufficient to observe that the system of critical orbits of $T_d \times \mathcal{T}$ remains the *same*, as in the case of T_d, and to study the action of reversal operations on the invariant spheres with stabilizers C_2 and C_s.

Internal stratification of the C_2-invariant spheres in the $\mathbb{C}P^2 \times \mathbb{S}^2$ space under the action of $T_d \times \mathcal{T}$ is shown in Figure 11, right. It can be seen that the action of $T_d \times \mathcal{T}$ on these spheres is equivalent to the natural action of the C_{2v} group (whose axis C_2 is the vertical axis in Figure 11). In addition to the two isolated S_4 points, this action creates one-dimensional strata with stabilizers $C_2 \times \mathcal{T}_2$ and $C_2 \times \mathcal{T}_s$.

The C_s-invariant spheres also do not remain homogeneous when reversal operations are accounted for properly. The action of $T_d \times \mathcal{T}$ on these spheres is equivalent to that of the group C_h (whose reflection plane is the plane of Figure 12); it creates a $(C_s \wedge \mathcal{T}_2)$-invariant circle on each C_s-invariant sphere in $\mathbb{C}P^2 \times \mathbb{S}^2$.

4.6. Action on the vibrational subspace $\mathbb{C}P^2 \times \mathbb{C}P^1$. Description of combined small-amplitude E- and F_2-mode vibrations requires the six-dimensional reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1$. The action of the $T_d \wedge \mathcal{T}$ symmetry group on the F_2-mode subspace $\mathbb{C}P^2$ and on the whole of $\mathbb{C}P^2 \times \mathbb{C}P^1$ is effective. The action of $T_d \times \mathcal{T}$ on the E-mode subspace is not effective; see Table 14: any point on this subspace is automatically D_2-invariant.

4.6.1. Critical orbits. All critical orbits on $\mathbb{C}P^2 \times \mathbb{C}P^1$ can be found and classified if we combine points from Tables 8 and 10 and determine the intersection of their stabilizers $G_f \cap G_e$ using the lattice from Figure 2, 3, or 8. To find whether the point (orbit) is critical, we consider the stratum with stabilizer $G_f \cap G_e$ and verify that we deal with an isolated point (orbit) using projections on the orbit spaces in Figures 9 and 10. As in the previous sections, we should combine points with identical stabilizers in order to determine the number of orbits and of points in the orbits. Concrete examples can be found in Table 15 if we ignore the rotational part (i.e., let $j = 0$). Results are summarized below.

Stabilizer of the orbits on			Number	Examples in
$\mathbb{C}P^2$	$\mathbb{C}P^1$	$\mathbb{C}P^2 \times \mathbb{C}P^1$	of points	Table 15
$D_{2d} \times \mathcal{T}$	$D_{2d} \times \mathcal{T}$	$D_{2d} \times \mathcal{T}$	$3+3$	$A_{1,2}^{(4)}$
$S_4 \wedge \mathcal{T}_s$	$D_{2d} \times \mathcal{T}$	$S_4 \wedge \mathcal{T}_s$	$6+6$	$(B,C)_{1,2}^{(4)}$
$C_{3v} \times \mathcal{T}$	$T \wedge \mathcal{T}_s$	$C_3 \wedge \mathcal{T}_s$	$4+4$	$A_{1,2}^{(3)}$
$C_3 \wedge \mathcal{T}_s$	$T \wedge \mathcal{T}_s$	$C_3 \wedge \mathcal{T}_s$	$8+8$	$(B,C)_{1,2}^{(3)}$
$C_{2v} \times \mathcal{T}$	$D_{2d} \times \mathcal{T}$	$C_{2v} \times \mathcal{T}$	$3+3$	$A_{1,2}^{(2)}$

Here the superscripts $^{(4)}$, $^{(3)}$, and $^{(2)}$ correspond to parts (a), (b), and (c) of Table 15; by $k+k$ we denote two k-point orbits.

4.6.2. Invariant subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1$. Three D_{2d}-invariant points on $\mathbb{C}P^2$ are the only points whose stabilizer includes D_2. Combining these points with the whole E-mode space $\mathbb{C}P^1 \sim \mathbb{S}^2$ gives three D_2-invariant spheres $\mathbb{S}^2$ in the six-dimensional space $\mathbb{C}P^2 \times \mathbb{S}^2$. Six C_s-invariant spheres on $\mathbb{C}P^2$ combined with, respectively, six D_{2d}-invariant points of the E-space give 12 C_s-invariant spheres $\mathbb{S}^2$ in $\mathbb{C}P^2 \times \mathbb{S}^2$. The stabilizer D_2 of the E-mode space $\mathbb{C}P^1 \sim \mathbb{S}^2$ has three conjugate C_2^a subgroups with $a = \{x, y, z\}$. Consequently, there are three C_2-invariant subspaces $\mathbb{S}^2_{(F)} \times \mathbb{S}^2_{(E)}$. Points with higher symmetry lie on each of the above subspaces.

4.7. Action on the full reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. In order to describe the simultaneous rotation and small amplitude E- and F_2-mode vibrations of the A_4 molecule (rotational structure of all combination polyads $n\nu_2 + m\nu_3$), we need the eight-dimensional total classical reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. In this section, we find critical (fixed) points of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ by combining points on the three factor spaces. Fortunately, most of the work has already been done in the previous sections. We can take critical orbits A, B, and C on $\mathbb{C}P^2 \times \mathbb{S}^2$ found in section 4.5 and combine them with critical orbits of the $T_d \times \mathcal{T}$ action on the E-mode space $\mathbb{C}P^1$ while matching the stabilizers carefully. It can be seen in the examples of Table 15 that orbits A, B, and C are duplicated in the presence of E-mode vibrations. Indexes 1 and 2 distinguish two different possible projections (z_5, z_6) on the E-mode space $\mathbb{C}P^1$. We use shorter T_d notation for stabilizers in Table 15. Extending the symmetry group T_d by the time reversal $\mathcal{T}$ does not modify the critical orbits; it merely doubles the order of stabilizers, which become $S_4 \wedge \mathcal{T}_s$, $C_3 \wedge \mathcal{T}_s$, and $C_s \times \mathcal{T}_2$. We comment briefly on these orbits.

The orbit stabilizer is defined as a class of conjugate subgroups; the point stabilizer is an individual subgroup in that class. As before, we match concrete point stabilizers. The S_4^z operation has $(1+2)$, 2, and 2 fixed points on the F_2-mode space $\mathbb{C}P^2$, E-mode space $\mathbb{C}P^1$, and rotational sphere $\mathbb{S}^2$, respectively. The $3 \times 2 \times 2 = 12$ points on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ with the stabilizer S_4^z are listed in Table 15. To complete orbits of the T_d group action we should add points with stabilizers S_4^x and S_4^y (which are obtained from the given points using symmetry operations R such that $S_4^\alpha = R \circ S_4^z \circ R^{-1}$). The total of 36 points can split into six orbits with stabilizer S_4. In particular, points A_1 and A_1' in Table 15 belong to the same six-point orbit (which also includes two points with stabilizer S_4^y and two points with stabilizer S_4^x).

A similar argument shows that there are 12 points on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ for each of the four conjugate stabilizers $C_3^{(k)}$ (Table 15(b), where points A', B', and C' are omitted for brevity). The total of $(3 \times 2 \times 2) \times 4 = 48$ points on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ is made up of six eight-point orbits with stabilizer C_3.

The remaining two critical orbits on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ have stabilizer C_s. We construct these orbits by taking points with stabilizer C_s^α on the $\mathbb{C}P^2 \times \mathbb{S}^2$ subspace (section 4.5) and combining them with the appropriate two D_{2d} points on the E-mode space $\mathbb{C}P^1$ (see Table 15(c)), as in the case of the stabilizer S_4. For six conjugate subgroups C_s we have 24 critical points on (the zero-dimensional component of) the C_s stratum of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, which separate into two 12-point orbits.

5. Prediction of RE. This section is a reward for our painstaking study of the $T_d \times \mathcal{T}$ group action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. The reduced Hamiltonian H_{eff} is a $(T_d \times \mathcal{T})$-invariant function on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ and it *must* have stationary points at all critical orbits of the $T_d \times \mathcal{T}$ action, which we found in the previous section (Table 15). Stationary points of H_{eff} are equilibria of the reduced system and are RE of the initial system with six vibrational degrees of freedom.[26]

[26] In order to simplify the analysis of the reduced system for P_4, the A_1-mode subsystem is excluded in [13] by setting $q^{A_1} = p^{A_1} = 0$ and $n_a = 0$; the model potential function was developed to cubic terms only. The number of degrees of freedom can be counted in several ways. Translational degrees of freedom can be trivially separated. Out of three rotational degrees of freedom, two can be reduced to account for the preservation of

RE correspond to families of special 3-tori in the initial phase space (see footnote 26) and are characterized *entirely* by symmetry and the values of the three integrals n_f, n_e, and j. The value of H_{eff} (energy) at RE and stability of RE are the primary characteristics of our system. Table 15 becomes, therefore, our most important result in view of practical applications. (Note that RE forming one symmetry group orbit are equivalent and it suffices to analyze stability and energy for one RE in the orbit.)

In this section, we are preoccupied with general understanding and description of possible functions H_{eff}. In particular, we want to know if such functions can have stationary points only on critical orbits, i.e., have the minimum number of stationary points. We place RE on critical group orbits and suggest possible stability. Assuming that H_{eff} is a Morse function on the space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, we make sure that possible sets of stationary points satisfy Morse inequalities for this space and its invariant subspaces. We then find the simplest possible set (or sets) of RE, which correspond to the simplest Morse Hamiltonian(s) on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ invariant with respect to the symmetry group of the system $T_d \times \mathcal{T}$. We can expect that at low perturbation (or excitation) our H_{eff} is such a simplest function.

The main purpose of this paper is the analysis of the complexity of the set of RE of systems combined of different subsystems. *Our first observation* in this context follows directly from the group action study in section 4. The RE set of the entire system is far from being a simple combination of the RE of the subsystems. As an example, take the F_2- and E-mode subsystems, which have 9 and 27 RE, respectively. These RE are also known as "nonlinear normal modes" [14, 15, 16]. They combine nonlinearly; there are 48 RE of the combined E–F_2 system corresponding to critical orbits on the $\mathbb{C}P^2 \times \mathbb{C}P^1$ phase space (see section 4.6).

Finding critical orbits on high-dimensional spaces constructed as a direct product of simpler spaces is greatly facilitated by tracing the correlation between critical orbits on subspaces and those on the complete space (for example, see section 4.5, Figures 11 and 12). *Our second main observation* is that the simplest $(T_d \times \mathcal{T})$-invariant Morse-type function H_{eff} with stationary points placed exclusively on critical orbits can be defined on individual subspaces $\mathbb{S}^2$, $\mathbb{C}P^1$, and $\mathbb{C}P^2$. However, when we go to a product space, such as $\mathbb{C}P^2 \times \mathbb{S}^2$, the situation becomes more complicated and there must be RE (stationary points of H_{eff}) lying on noncritical group orbits.

5.1. Consequences of local symmetry for linear stability. We can use our results on the critical orbits of the $T_d \times \mathcal{T}$ group action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ to find the position of corresponding RE and compute their energy for any given Hamiltonian H_{eff}. To find more about these RE *without* using any concrete H_{eff}, we should classify small phase space displacements x from them (coordinates on the tangent plane) according to the irreducible representations of their stabilizers. In this work, we are interested in one particular kind of group action, which

the length of the angular momentum vector and its projection on a laboratory fixed frame. This leaves six vibrational degrees of freedom and one rotational degrees of freedom, sometimes called "internal" degrees of freedom. However, we replace internal rotational degrees of freedom with two constrained oscillatory degrees of freedom, so there is a total of eight initial degrees of freedom (and a phase space of dimension 16), of which two represent one physical rotational degrees of freedom, and one representing the A_1 mode is ignored. Reduction with respect to all integrals n_a, n_f, n_e, and j leads to the reduced system on the space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ with four degrees of freedom. The corresponding dynamical symmetry is T^4. If we neglect n_a by setting $q^{A_1} = p^{A_1} = 0$ *before* reduction, we restrict the initial system to seven degrees of freedom, and the dynamical symmetry is T^3.

is induced on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ by spatial and reversal symmetries of the initial molecular rotation–vibration system. This action can be studied in the example of the SO(2) group (axial symmetry) and its discrete cyclic subgroups C_k (rotation by $2\pi/k$ with $k = 2, 3, 4, \dots$) acting on the initial coordinates of our system as groups of transformations.

We begin by describing a number of nondegenerate equilibria e with nontrivial stabilizers $G_e = C_k$ and corresponding G_e-invariant local quadratic Hamiltonians H_0 such that $H_0(e) = 0$. The local symmetry of e results in certain restrictions on H_0 and on the stability. We compute the Hessian matrix $\partial^2 H_0/\partial x^2$. Displacements that transform according to different rows of the same irreducible representation Γ_e form a degenerate eigenvalue block of $\partial^2 H_0/\partial x^2$. The number of negative eigenvalues of this matrix, or the *Morse index*, enters in Morse inequalities in section 5.2; the *Poincaré index* gives the sign of $\det(\partial^2 H_0/\partial x^2)$. In order to determine the stability of e, we consider H_0 as a Hamiltonian, establish local symplectic structure $\mathcal{J}$, and compute the eigenvalues of the corresponding Hamiltonian matrix $\mathcal{J}\partial^2 H_0/\partial x^2$. Hamiltonian stability and the Morse index are related unambiguously only in the case of dimension two (reduced system with one degree of freedom). The Poincaré index, which can be computed from the Morse index, gives some characteristics of linear Hamiltonian stability in systems with any degrees of freedom. Subsequently, we use our results to study the RE of our system.

5.1.1. C_k-invariant quadratic Hamiltonians on $\mathbb{R}^2$. We define the action of the rotation group SO(2) on the real variables (x_1, x_2) using the 2×2 orthogonal matrix

$$M_m(\varphi) = \begin{pmatrix} \cos m\varphi & \sin m\varphi \\ -\sin m\varphi & \cos m\varphi \end{pmatrix}, \quad m = 0, 1, 2, \dots,$$

and consider quadratic functions $H_0(x_1, x_2)$, which are invariant with respect to SO(2) or its discrete subgroups C_k (rotation by $\varphi = 2\pi/k$). For each given m we take $k > m$. Furthermore, it suffices to consider $k \leq 2m$. The point $e = (x_1, x_2) = 0$ is a critical orbit and, therefore, an equilibrium of H_0. If the two displacements (x_1, x_2) about e transform as components of a two-dimensional real irreducible representation, then e can only be a minimum or a maximum of H_0 with signature $(++)$ or $(--)$ (Morse index 0 or 2), respectively, but not a saddle $(+-)$ with index 1.

Invariants of our SO(2) action are constructed using combinations

$$\xi = x_1 + ix_2 \quad \text{and} \quad \bar{\xi} = x_1 - ix_2,$$

which transform as conjugate irreducible one-dimensional complex representations $\pm m$ of SO(2). The pair (x_1, x_2) realizes a two-dimensional representation of SO(2), which is *irreducible over reals*. Descending to C_k, we should check whether this representation remains irreducible (over reals). For $m = 1$ and $m = 2$ this is the case for all $k \neq 2$ and $k \neq 4$, respectively. When $m = 2$ and $k = 4$, the image of C_4 becomes C_2. The corresponding quadratic forms are given in Table 16. We can see that the case $k = 2m$ is stable (elliptic) when $c^2 - ab < 0$ or unstable (hyperbolic) otherwise, while the $k \neq 2m$ equilibrium is always stable.

Table 16

Invariant local quadratic Hamiltonians with one and two degrees of freedom encountered in our study. The action of C_k on the phase plane $\mathbb{R}^2$ with coordinates (x_1, x_2) is given by the 2×2 orthogonal matrix $M_m(\varphi)$ with $\varphi = 2\pi/k$; the action of C_k on the phase space $\mathbb{R}^4$ with coordinates (x_1, x_2, x_3, x_4) is defined by $\operatorname{diag}(M_{m'}, M_{m''})(\varphi)$*. The action of the subgroups of $T_d \times \mathcal{T}$ on the local coordinates (x_1, x_2, y_1, y_2) is given in Table* 18.

Stabilizer	C_k-invariant quadratic form	Eigenvalues[27,28]
Case 2D: $m = 1, 2$, etc., $k > m$, and $\omega = dx \wedge dy$		
C_{2m}	$\frac{1}{2}ax^2 + \frac{1}{2}by^2 + cxy$	$\left[\pm\sqrt{c^2 - ab}\right], \quad \frac{1}{2}\{a + b \pm D_{abc}\}$
C_{2m+1}	$\frac{1}{2}a(x^2 + y^2)$	$[\pm ia], \quad \{a, a\}$
Case A: $m' = 1$, $m'' = 1$, and $\omega = dx_1 \wedge dy_1 + dx_2 \wedge dy_2$		
C_2	$\frac{1}{2}a'x_1^2 + \frac{1}{2}a''x_2^2 + \frac{1}{2}b'y_1^2 + \frac{1}{2}b''y_2^2$ + (all possible cross terms)	No restrictions
$C_k (k > 2)$	$\frac{1}{2}a(x_1^2 + x_2^2) + \frac{1}{2}b(y_1^2 + y_2^2)$ $+ c(x_1y_1 + x_2y_2) + d(x_1y_2 - x_2y_1)$	$\left[\pm\sqrt{c^2 - ab} + id, \pm\sqrt{c^2 - ab} - id\right]$ $\frac{1}{2}\left\{a + b \pm D_{abcd}\right\} \times 2$
Case A with extended symmetry $\mathcal{T}_k = C_k \times \mathcal{T}$, where $\mathcal{T} : (x_1, x_2, y_1, y_2) \to (x_1, x_2, -y_1, -y_2)$		
$\mathcal{T}_2$	$\frac{1}{2}a'x_1^2 + \frac{1}{2}a''x_2^2 + \frac{1}{2}b'y_1^2 + \frac{1}{2}b''y_2^2$ $+ cx_1x_2 + dy_1y_2$	No simple restrictions $\frac{1}{2}\left\{b' + b'' \pm D_{b'b''d}, a' + a'' \pm D_{a'a''c}\right\}$
$\mathcal{T}_k (k > 2)$	$\frac{1}{2}a(x_1^2 + x_2^2) + \frac{1}{2}b(y_1^2 + y_2^2)$	$[\pm i\sqrt{ab}] \times 2, \quad \{a, b\} \times 2$
Case A with full symmetry and $\omega = dx_1 \wedge dy_1 + dx_2 \wedge dy_2$		
$C_{2v} \times \mathcal{T}$	$\frac{1}{2}a'x_1^2 + \frac{1}{2}a''x_2^2 + \frac{1}{2}b'y_1^2 + \frac{1}{2}b''y_2^2$	$\left[\pm i\sqrt{a'b'}, \pm i\sqrt{a''b''}\right], \quad a', b', a'', b''$
$D_{2d} \times \mathcal{T}$	$\frac{1}{2}a(x_1^2 + x_2^2) + \frac{1}{2}b(y_1^2 + y_2^2)$	$[\pm i\sqrt{ab}] \times 2, \quad \{a, b\} \times 2$
$C_{3v} \times \mathcal{T}$	$\frac{1}{2}a(x_1^2 + x_2^2) + \frac{1}{2}b(y_1^2 + y_2^2)$	$[\pm i\sqrt{ab}] \times 2, \quad \{a, b\} \times 2$
Case B: $m' = 1$, $m'' = 2$, and $\omega = dx_1 \wedge dy_1 + dx_2 \wedge dy_2$		
C_3	$\frac{1}{2}a'(x_1^2 + y_1^2) + \frac{1}{2}a''(x_2^2 + y_2^2)$ $+ c(x_1x_2 - y_1y_2) + d(x_1y_2 + y_1x_2)$	$\frac{1}{2}\left[i(a' - a'') \pm \Delta, i(a'' - a') \pm \Delta\right]$ $\frac{1}{2}\left\{a' + a'' \pm D_{a'a''cd}\right\} \times 2$
C_4	$\frac{1}{2}a'(x_1^2 + y_1^2) + \frac{1}{2}a''x_2^2 + \frac{1}{2}b''y_2^2 + cx_2y_2$	$[\pm ia', \pm\sqrt{c^2 - a''b''}]$ $\left\{a', a', \frac{1}{2}(a'' + b'' \pm D_{a''b''c})\right\}$
$C_k (k > 4)$	$\frac{1}{2}a'(x_1^2 + y_1^2) + \frac{1}{2}a''(x_2^2 + y_2^2)$	$[\pm ia', \pm ia''], \quad \{a', a''\} \times 2$
Case B with full symmetry and $\omega = dx_1 \wedge dy_1 + dx_2 \wedge dy_2$		
$C_3 \wedge \mathcal{T}_s$	$\frac{1}{2}a'(x_1^2 + y_1^2) + \frac{1}{2}a''(x_2^2 + y_2^2)$ $+ c(x_1x_2 - y_1y_2)$	$\frac{1}{2}\left[i(a' - a'') \pm \Delta, i(a'' - a') \pm \Delta\right]$ $\frac{1}{2}\left\{a' + a'' \pm D_{a'a''c}\right\} \times 2$
$S_4 \wedge \mathcal{T}_2$	$\frac{1}{2}a'(x_1^2 + y_1^2) + \frac{1}{2}a''x_2^2 + \frac{1}{2}b''y_2^2$	$\left[\pm ia', \pm i\sqrt{a''b''}\right], \quad \{a', a', a''b''\}$

5.1.2. C_k-invariant quadratic Hamiltonians on $\mathbb{R}^4$.

We define the action of the SO(2) group on the four-plane $\mathbb{R}^4$ with coordinates (x_1, x_2, x_3, x_4) using the matrix $\left(\begin{smallmatrix} M_{m'} & 0 \\ 0 & M_{m''} \end{smallmatrix}\right)$, where the submatrices $M_{m'}$ and $M_{m''}$ act on the (x_1, x_2)-subspace and the (x_3, x_4)-subspace,

[27] Eigenvalues of the Hamiltonian and Hessian matrices are given in square [] and curly {} brackets, respectively; ×2 indicates multiplicity.

[28] $\Delta = \sqrt{4(c^2 + d^2) - (a' + a'')^2}$, $D_{abcd} = \sqrt{(a - b)^2 + 4(c^2 + d^2)}$, $D_{abc} = D_{abc0}$.

respectively. (The case of the diagonal SO(2) action on $\mathbb{R}^2 \times \mathbb{R}^2$.) Invariants of this action can be readily constructed using

$$\xi = (x_1 + ix_2) \quad \text{and} \quad \eta = (x_3 + ix_4)$$

together with their conjugates $\bar{\xi}$ and $\bar{\eta}$. The four variables ξ, η, $\bar{\xi}$, $\bar{\eta}$ realize irreducible representations m', m'', $-m'$, $-m''$, respectively. We consider several situations of interest to our later study.

In the case of $m' = 1$ and $m'' = 2$, we have two quadratic SO(2) invariants $\frac{1}{2}\xi\bar{\xi}$ and $\frac{1}{2}\eta\bar{\eta}$. When SO(2) is lowered to C_3, we also have $\xi\eta$ and $\bar{\xi}\bar{\eta}$, which transform like $\exp(\pm 3i\varphi)$. Similarly, η^2 and $\bar{\eta}^2$ transform like $\exp(\pm 4i\varphi)$ and are the two extra invariants in the case of C_4.

In the case of $m' = m'' = 1$ (and generally for $m' = m''$), our SO(2) action has four quadratic invariants: the familiar $\frac{1}{2}\xi\bar{\xi}$, $\frac{1}{2}\eta\bar{\eta}$, and the cross terms $\xi\bar{\eta}$, $\eta\bar{\xi}$. The same four remain if SO(2) is lowered to C_k and $k > 2$. When $k = 2$, each coordinate x_i realizes real one-dimensional irreducible antisymmetric representation. All 10 quadratic monomials $x_j x_i$ are, therefore, C_2-invariant.

Generic C_k-invariant real quadratic forms in (x_1, x_2, x_3, x_4) constructed using the above invariants are presented in Table 16. As can be seen from this table, Hamiltonian stability of $e = (0, 0, 0, 0)$ depends on m', m'', k, and the symplectic form ω. The eigenvalues of the Hessian at $(0, 0, 0, 0)$, which are also given in Table 16, are used in the Morse theory analysis. Furthermore, we can see that additional symmetry, such as the time reversal extension for case A critical orbits (see sections 4.3, 4.5.3 and Table 15), can simplify the situation quite radically.

5.1.3. Points on $\mathbb{S}^2$ and $\mathbb{C}P^1$. Consider a nondegenerate critical point e with stabilizer G_e on the 2-sphere $\mathbb{S}^2$ or on the diffeomorphic space $\mathbb{C}P^1$. The group G_e is defined as a group of transformations of the ambient Euclidean space $\mathbb{R}^3$, which embeds $\mathbb{S}^2$ (see section 4). We want to know how G_e acts on the 2-plane $\mathbb{R}^2_{(e)}$ tangent to $\mathbb{S}^2$ or $\mathbb{C}P^1$ at e. It suffices to consider the circle group $G_e = \mathrm{SO}(2)$ and its subgroups C_k. A straightforward computation shows that the image of SO(2) and C_k in the representation spanned by the Euclidean coordinates (x, y) on $\mathbb{R}^2_{(e)}$ is again SO(2) and C_k and that $x \pm iy$ span representations ± 1. Then, following section 5.1.1, the point e on $\mathbb{S}^2$ and $\mathbb{C}P^1$ with stabilizer C_k, $k > 2$, is *always* stable; points with stabilizer C_2 can also be unstable.

5.1.4. Points on $\mathbb{C}P^2$. As before, we study a special case of the group action on the $\mathbb{C}P^2$ space induced by the natural action (vector representation) of the group SO(2) and its subgroups C_k on the complex 3-space with coordinates (z_1, z_2, z_3). This action is defined by a 3×3 orthogonal matrix M. The action on the corresponding real 6-space with coordinates $(q_1, q_2, q_3, p_1, p_2, p_3)$ is given by the matrix $\left(\begin{smallmatrix} M & 0 \\ 0 & M \end{smallmatrix}\right)$ with one copy of M acting on the q space and the other on the p space. The above action of SO(2) with the symmetry axis z_3, and of the corresponding discrete subgroups C_k with $k > 2$, on $\mathbb{C}P^2$ has three isolated fixed points (see section 4.3 and [17]):

$$A = (0, 0, 1), \qquad B = (1, \pm i, 0).$$

Table 17

Generic equilibria on $\mathbb{C}P^2$ in the presence of the natural action of the cyclic group C_k and its extensions.

Equilibrium[29]		$\partial^2 H_0$[30]	$\mathcal{J}\partial^2 H_0$[31]	
A	$C_2 \times \mathcal{T}$	No simple restrictions		
	$C_{2v} \times \mathcal{T}$	4, 2, 0	ii	
		3, 1	ir	
		2	rr	
A	$C_k \times \mathcal{T},\ k > 2$	4, 0	ii	1:1
		2	rr	1:1

Equilibrium[29]		$\partial^2 H_0$[30]	$\mathcal{J}\partial^2 H_0$[31]
B	C_3	4, 2, 0	ii
		2	c
B	C_4	4, 2, 0	ii
		3, 1	ir
B	$C_k,\ k > 4$	4, 2, 0	ii

(In Table 15 we use notation B, C for points of type B.) In the C_2 case, only the A-type fixed point is isolated. It is also useful to recall that the stabilizer of the A points in the case of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^2$ includes $\mathcal{T}$ (see Table 10) and that C_k can be extended easily to $\mathcal{T}_k = C_k \times \mathcal{T}$.

Any C_k-invariant Morse Hamiltonian H on $\mathbb{C}P^2$ has nondegenerate stationary points of types A and B. Let z be one of these points and let $\mathbb{C}^2_{(z)} \sim \mathbb{R}^4_{(z)}$ be the plane tangent to $\mathbb{C}P^2$ at z. This plane is a chart of $\mathbb{C}P^2$ with four real displacement coordinates (q', p', q'', p''). The zero order $H_0(q', p', q'', p'')$ of the Taylor expansion of H near z is a nondegenerate quadratic form. We study the action of C_k on (q', p', q'', p'') and find which generic forms in Table 16 correspond to H_0. Table 17 gives the summary of the results.

The action of C_k on $\mathbb{R}^4_{(z)}$ can be found by direct computation; see also section 10. The matrix of the rotation about axis z_3 by an arbitrary angle φ is

$$M(\varphi) = \begin{pmatrix} \cos\varphi & \sin\varphi & 0 \\ -\sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We can see right away that four small real quantities (q_1, q_2), and (p_1, p_2), which define a displacement from point A (in the appropriate chart of $\mathbb{C}P^2$),

$$d_A = (q_1 + ip_1, q_2 + ip_2, 1)$$

transform like two copies of the real two-dimensional representation ±1 of the group SO(2). Consequently, the forms of case A in Table 16 with (x_1, x_2, y_1, y_2) corresponding to (q_1, q_2, p_1, p_2) represent generic C_k and $(C_k \times \mathcal{T})$-invariant Morse functions locally at point A. The $C_2 \times \mathcal{T}$ form can be further simplified if we consider the full $C_{2v} \times \mathcal{T}$ stabilizer of the actual fixed point of the $T_d \times \mathcal{T}$ action. In that case H_0 has only four terms q_1^2, q_2^2, p_1^2, and p_2^2. Hamiltonian (linear) stability analysis of the A-type equilibria in the case of the full $T_d \times \mathcal{T}$ action is straightforward (see Table 17) because the linearized system separates in initial phase space coordinates and the analysis reduces to combining two systems with one degree of freedom.

[29] Type and local symmetry (stabilizer) of equilibria on $\mathbb{C}P^2$; see section 5.1.4 and Table 15 and compare to Table 16.

[30] Possible Morse index derived from the corresponding Hessian eigenvalues in Table 16.

[31] Possible eigenvalues of the Hamiltonian matrix: i and r stand for the imaginary and the real eigenvalue pair, respectively; 1:1 indicates resonance, c denotes four complex eigenvalues.

When we use the same approach for

$$d_B = (1, i + q_2 + ip_2, q_3 + ip_3) = (1, i + z_2, z_3),$$

we should project the transformed vector

$$M(\varphi)d_B = \left(e^{i\varphi} + z_2 \sin\varphi,\ ie^{i\varphi} + z_2\cos\varphi,\ z_3\right)$$

back to the initial chart of $\mathbb{C}P^2$ and Taylor expand to the first order in (q_2, p_2, q_3, p_3),

$$\begin{aligned} M(\varphi)d_B\big|_{\text{chart}} &= \left[e^{i\varphi} + z_2 \sin\varphi\right]^{-1} M(\varphi)d_B \\ &\approx \left(1,\ i + e^{-2i\varphi}z_2,\ e^{-i\varphi}z_3\right). \end{aligned}$$

We can now see that (q_3, p_3) and (q_2, p_2) realize representations ±1 and ±2 of SO(2), respectively. Therefore, quadratic forms of case B in Table 16 with (x_1, y_1, x_2, y_2) corresponding to (q_2, p_2, q_3, p_3) represent generic C_k-invariant Morse functions locally at point B. For all points $B^{(k)}$ with $k > 3$, linearization separates in the initial coordinates (q_2, p_2) and (q_3, p_3). As in case A, stability analysis of these equilibria is simple. Point $B^{(3)}$ turns out to be the only interesting case, where the localized system is intrinsically four-dimensional; cf. [97].

5.1.5. Stability analysis of stationary points on $\mathbb{C}P^2$ in the presence of $\mathcal{T}_d \times \mathcal{T}$. In the previous section, we showed that linear stability of stationary points on $\mathbb{C}P^2$ (vibrational RE) can be predicted by analyzing possible local Hamiltonians for RE whose stabilizer is an SO(2) group or a discrete cyclic subgroup C_k of this group. For the five types of vibrational RE labeled $A^{(2)}$, $A^{(3)}$, $B^{(3)}$, $A^{(4)}$, and $B^{(4)}$ (see Table 15) we take subgroups C_2, C_3, and C_4, respectively. The latter can be regarded as the principal symmetry operations of the respective stabilizers $C_{2v} \times \mathcal{T}$, $C_{3v} \times \mathcal{T}$, $C_3 \wedge \mathcal{T}$, $D_{2d} \times \mathcal{T}$, and $S_4 \wedge \mathcal{T}$. At the same time, prediction of stability of these RE can be further improved if we account for the full stabilizers.

To this end we proceed as before in section 5.1.4. We define local displacement coordinates (x_1, y_1, x_2, y_2) near the stationary point on $\mathbb{C}P^2$ and determine the action of the full stabilizer G on these coordinates. Knowing the action, we find the representation of G realized by (x_1, y_1, x_2, y_2) and construct the typical G-invariant quadratic form $H(x_1, y_1, x_2, y_2)$. We choose (x_1, y_1, x_2, y_2) so that the local 2-form is $dx_1 \wedge dy_1 + dx_2 \wedge dy_2$ and consider $H(x_1, y_1, x_2, y_2)$ as a Hamiltonian function of local linearization near the RE.

In order to define (x_1, y_1, x_2, y_2) we rotate the initial coordinates (z_1, z_2, z_3) so that in the new coordinates (z'_1, z'_2, z'_3) the principal symmetry axis of the stabilizer becomes axis z'_1. The direction of the two other axes can be chosen as shown below:

Stabilizer	(z_1, z_2, z_3)	Map	(z'_1, z'_2, z'_3)
$D_{2d}^{(x)} \times \mathcal{T}$	$\sqrt{2n}(1, 0, 0)$	E	$\sqrt{2n}(1, 0, 0)$
$S_4^{(x)} \wedge \mathcal{T}_2^{(y)}$	$\sqrt{n}(0, 1, \mathrm{i})$	E	$\sqrt{n}(0, 1, i)$
$C_{2v}^{(z)} \times \mathcal{T}$	$\sqrt{n}(1, 1, 0)$	M_1	$\sqrt{2n}(1, 0, 0)$
$C_{3v}^{[111]} \times \mathcal{T}$	$\sqrt{\frac{2}{3}n}(1, 1, 1)$	M_2	$\sqrt{2n}(1, 0, 0)$
$C_3^{[111]} \wedge \mathcal{T}_s^{\Vert}$	$\sqrt{\frac{2}{3}n}(1, \eta^2, \bar{\eta}^2)$	M_2	$\sqrt{n}e^{\mathrm{i}\pi/6}(0, 1, i)$

where $z' = Mz$, $E = \operatorname{diag}(1, 1, 1)$,

$$M_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{and} \quad M_2 = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{pmatrix}.$$

Using the approach in section 5.1.4, we can show that the position of the RE in the new coordinates (z'_1, z'_2, z'_3) is $(1,0,0)$ for type A and $(0,1,\pm i)$ for type B. We define displacement vectors

$$d_A = (z_1, x_1 + iy_1, x_2 + iy_2),$$

with $z_1 = \sqrt{2n - |z_2|^2 - |z_3|^2}$, and

$$d_B = (x_1 + iy_1, z'_2, i\sqrt{n} + \sqrt{2}x_2 + iy_2/\sqrt{2}),$$

with $z'_2 = \sqrt{n - |z'_1|^2 - |z'_3|^2}$. Note that, instead of projecting on an $\mathbb{R}^4_{(x,y)}$ chart of $\mathbb{C}P^2$ as we do in section 5.1.4, we represent the points of $\mathbb{C}P^2$ by fixing the total phase of (z'_1, z'_2, z'_3) so that $\mathrm{Im}(z'_1) = 0$ in the case of d_A and $\mathrm{Im}(z'_2) = 0$ in the case of d_B (cf. section 2.4.3). Each symmetry operation R in the stabilizer G is realized initially as a linear transformation of the space $\mathbb{R}^3$ defined by a real matrix M_R. The same transformation applies to $\mathbb{C}^3$ with coordinates (z'_1, z'_2, z'_3). To realize this transformation on $\mathbb{C}P^2$, we should *correct* or *restore* the phase of $M_R z'$ in order to obey our phase condition. Therefore we define

$$Rd_A = \frac{[M_R \bar{z}]_1}{|[M_R \bar{z}]_1|} M_R d_A \quad \text{and} \quad Rd_B = \frac{[M_R \bar{z}]_2}{|[M_R \bar{z}]_2|} M_R d_B.$$

To find the action on (x,y) we Taylor expand Rd_A and Rd_B at $(x,y) = 0$ and compare them to the initial vectors d_A and d_B. Results are summarized in Table 18. As can be concluded from this table, displacements (x,y) realize the following representations of the respective stabilizers:

Point	Stabilizer	Representation spanned by (x,y)
$A^{(2)}$	$C_{2v} \times \mathcal{T}$	$A_{2g}(x_1) \oplus A_{2u}(y_1) \oplus B_{1g}(x_2) \oplus B_{1u}(y_2)$
$A^{(3)}$	$C_{3v} \times \mathcal{T}$	$E_g(x_1, x_2) \oplus E_u(y_1, y_2)$
$B^{(3)}$	$C_3 \wedge \mathcal{T}_s$	$E(x_1, y_1) \oplus E(x_2, -y_2)$
$A^{(4)}$	$D_{2d} \times \mathcal{T}$	$E_g(x_1, x_2) \oplus E_u(y_1, y_2)$
$B^{(4)}$	$S_4 \wedge \mathcal{T}_2$	$B_1(y_2) \oplus B_2(x_2) \oplus E(x_1, y_1)$

Here we denote irreducible representations of $\mathcal{T}$-extended groups using notation of corresponding point groups [66].

The above decomposition of representations realized by local displacements into irreducible representations makes construction of local quadratic Hamiltonians straightforward. In the case of $A^{(k)}$ and $B^{(4)}$ RE, all quadratic invariants are just linear combinations of scalar squares of displacements transforming according to different irreducible representations, such as $\left[E_g(x_1, x_2)\right]^2 = x_1^2 + x_2^2$, etc. The case of $B^{(3)}$ is the only case where a scalar product of two different displacements transforming according to the same irreducible representation E occurs. Generic local quadratic Hamiltonians for each RE are listed in Table 16. The brute-force way to find these Hamiltonians is by projecting the most general homogeneous second degree polynomial in (x_1, x_2, y_1, y_2) using the operator $|G|^{-1} \sum_{R \in G} R$, where G is the stabilizer of the RE in question.

Table 18

Action of stabilizers on local displacements from the stationary points on $\mathbb{C}P^2$.

Action of $D_{2d}^{(x)} \times \mathcal{T}$ on $E_g \oplus E_u$

R	Rx_1	Rx_2	Ry_1	Ry_2	R	Rx_1	Rx_2	Ry_1	Ry_2
E	x_1	x_2	y_1	y_2	$\mathcal{T}$	x_1	x_2	$-y_1$	$-y_2$
C_2^x	$-x_1$	$-x_2$	$-y_1$	$-y_2$	$\mathcal{T}_2^x$	$-x_1$	$-x_2$	y_1	y_2
C_2^y	$-x_1$	x_2	$-y_1$	y_2	$\mathcal{T}_2^y$	$-x_1$	x_2	y_1	$-y_2$
C_2^z	x_1	$-x_2$	y_1	$-y_2$	$\mathcal{T}_2^z$	x_1	$-x_2$	$-y_1$	y_2
σ^{yz}	x_2	x_1	y_2	y_1	$\mathcal{T}_s^{yz}$	x_2	x_1	$-y_2$	$-y_1$
$\sigma^{\overline{yz}}$	$-x_2$	$-x_1$	$-y_2$	$-y_1$	$\mathcal{T}_s^{\overline{yz}}$	$-x_2$	$-x_1$	y_2	y_1
S_4	x_2	$-x_1$	y_2	$-y_1$	$S_4\mathcal{T}$	x_2	$-x_1$	$-y_2$	y_1
S_4^{-1}	$-x_2$	x_1	$-y_2$	y_1	$S_4^{-1}\mathcal{T}$	$-x_2$	x_1	y_2	$-y_1$

Action of $C_{2v}^{(z)} \times \mathcal{T}$ on $A_{2g} \oplus A_{2u} \oplus B_{1g} \oplus B_{1u}$

R	Rx_1	Rx_2	Ry_1	Ry_2	R	Rx_1	Rx_2	Ry_1	Ry_2
E	x_1	x_2	y_1	y_2	$\mathcal{T}$	x_1	x_2	$-y_1$	$-y_2$
C_2^z	x_1	$-x_2$	y_1	$-y_2$	$\mathcal{T}_2^z$	x_1	$-x_2$	$-y_1$	y_2
σ^{xy}	$-x_1$	x_2	$-y_1$	y_2	$\mathcal{T}_s^{xy}$	$-x_1$	x_2	y_1	$-y_2$
$\sigma^{\overline{xy}}$	$-x_1$	$-x_2$	$-y_1$	$-y_2$	$\mathcal{T}_s^{\overline{xy}}$	$-x_1$	$-x_2$	y_1	y_2

Action[32] of $C_{3v}^{[111]} \times \mathcal{T}$ on $E_g \oplus E_u$

R	Rx_1	Rx_2	Ry_1	Ry_2
E	x_1	x_2	y_1	y_2
C_3	$-ax_1 - bx_2$	$bx_1 - ax_2$	$-ay_1 - by_2$	$by_1 - ay_2$
C_3^2	$-ax_1 + bx_2$	$-bx_1 - ax_2$	$-ay_1 + by_2$	$-by_1 - ay_2$
σ_{xy}	x_1	$-x_2$	y_1	$-y_2$
σ_{yz}	$-ax_1 + bx_2$	$bx_1 + ax_2$	$-ay_1 + by_2$	$by_1 + ay_2$
σ_{zx}	$-ax_1 - bx_2$	$-bx_1 + ax_2$	$-ay_1 - by_2$	$-by_1 + ay_2$
$\mathcal{T}$	x_1	x_2	$-y_1$	$-y_2$
$C_3\mathcal{T}$	$-ax_1 - bx_2$	$bx_1 - ax_2$	$ay_1 + by_2$	$-by_1 + ay_2$
$C_3^2\mathcal{T}$	$-ax_1 + bx_2$	$-bx_1 - ax_2$	$ay_1 - by_2$	$by_1 + ay_2$
$\mathcal{T}_s^{xy}$	x_1	$-x_2$	$-y_1$	y_2
$\mathcal{T}_s^{yz}$	$-ax_1 + bx_2$	$bx_1 + ax_2$	$ay_1 - by_2$	$-by_1 - ay_2$
$\mathcal{T}_s^{zx}$	$-ax_1 - bx_2$	$-bx_1 + ax_2$	$ay_1 + by_2$	$by_1 - ay_2$

Action of $S_4^{(x)} \wedge \mathcal{T}_2^{(y)}$ on $B_1 \oplus B_2 \oplus E$

R	Rx_1	Rx_2	Ry_1	Ry_2	R	Rx_1	Rx_2	Ry_1	Ry_2
E	x_1	x_2	y_1	y_2	$\mathcal{T}_2^y$	$-x_1$	$-x_2$	y_1	y_2
C_2^x	$-x_1$	x_2	$-y_1$	y_2	$\mathcal{T}_2^z$	x_1	$-x_2$	$-y_1$	y_2
S_4	y_1	$-x_2$	$-x_1$	$-y_2$	$\mathcal{T}_s^{yz}$	y_1	x_2	x_1	$-y_2$
S_4^{-1}	$-y_1$	$-x_2$	x_1	$-y_2$	$\mathcal{T}_s^{\overline{yz}}$	$-y_1$	x_2	$-x_1$	$-y_2$

Action[32] of $C_3^{[111]} \wedge \mathcal{T}_s^{\|}$ on $E \oplus E$

R	Rx_1	Rx_2	Ry_1	Ry_2
E	x_1	x_2	y_1	y_2
C_3	$-ax_1 - by_1$	$-ax_2 + by_2$	$bx_1 - ay_1$	$-bx_2 - ay_2$
C_3^2	$-ax_1 + by_1$	$-by_2 - ax_2$	$-bx_1 - ay_1$	$bx_2 - ay_2$
$\mathcal{T}_s^{xy}$	$ax_1 - by_1$	$ax_2 + by_2$	$-bx_1 - ay_1$	$bx_2 - ay_2$
$\mathcal{T}_s^{yz}$	$-x_1$	$-x_2$	y_1	y_2
$\mathcal{T}_s^{zx}$	$ax_1 + by_1$	$ax_2 - by_2$	$bx_1 - ay_1$	$-bx_2 - ay_2$

[32] Notation $a = 1/2$, $b = \sqrt{3}/2$.

5.2. Application of Morse theory. Simplest Morse Hamiltonians. Consider a manifold P whose topology is described by $\dim P + 1$ Betti numbers b_k. Particularly useful is the combination of these numbers, called the Euler characteristics Σ. A Morse function f on P is smooth and has only nondegenerate stationary points. Let c_k be the number of stationary points of f of Morse index k. The set of $\dim P$ Morse inequalities

$$\sum_{k=0}^{s}(-1)^{s-k}c_k \geq \sum_{k=0}^{s}(-1)^{s-k}b_k, \quad 0 \leq s < \dim P,$$

and the Euler–Poincaré equation

$$\sum_{k=0}^{\dim P}(-1)^{k}\,c_k = \sum_{k=0}^{\dim P}(-1)^{k}\,b_k = \Sigma,$$

express the relation between c_k and topological invariants b_k and Σ.

In the presence of a nonfree action of group G on P, all isolated points on the critical orbits of this action must be stationary points of f. The Morse function f with a *minimal possible* number of stationary points on P (in the presence of the specific group action) represents a class of simplest Morse functions. In the most trivial situations, such functions would have stationary points only on the isolated critical points. We should, therefore, check whether (and how) placing stationary points exclusively on the isolated points of critical orbits can satisfy the above Morse theory requirements. Table 19 gives Betti numbers for $\mathbb{S}^2$ and $\mathbb{C}P^2$ and suggests systems of stationary points on the vibrational spaces $\mathbb{C}P^2$ and $\mathbb{C}P^1$ and the rotational space $\mathbb{S}^2$ satisfying Morse theory in the presence of the $T_d \times \mathcal{T}$ group action. We begin with the simplest Morse Hamiltonians on each factor space of the total reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. Such Hamiltonians describe isolated F_2-mode or E-mode vibrational systems (polyads) or pure rotation.

5.2.1. Morse functions on the rotational space $\mathbb{S}^2$. Among the 26 fixed points of the $T_d \times \mathcal{T}$ action on $\mathbb{S}^2$ (Table 7), six points with stabilizer $S_4 \wedge \mathcal{T}$ and eight points with stabilizer $C_3 \wedge \mathcal{T}_s$ should be elliptic. The Morse conditions are satisfied if the 12 points with stabilizer $C_s \times \mathcal{T}_2$ are hyperbolic (unstable): $6 - 12 + 8 = 2$. The two possible simplest Morse Hamiltonians differ in sign: one has six maxima and eight minima while the other has this structure turned upside-down. If the internuclear adiabatic potential of the A_4 molecule can be well approximated as a sum of six pairwise interaction terms and all vibrations are frozen, then the minima are located at the six S_4 points [12] as shown in Figure 13, left.

We like to note that stationary points of simplest Morse functions of purely rotational systems (rotational RE of nonrigid bodies) should not necessarily be fixed points on $\mathbb{S}^2$. Thus in the case of the lowest possible symmetry of such systems $\mathcal{T}$ (no spatial symmetry), neither of the three pairs of equivalent RE has a fixed position on $\mathbb{S}^2$. Another example is the $C_2 \times \mathcal{T}$ system in Figure 6: four (two pairs) of its six RE can lie anywhere on the invariant circle.

5.2.2. Morse functions on the E-mode phase space $\mathbb{C}P^1$. Critical orbits of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^1$ are presented in Figure 9 and Table 8. The two equivalent $T \wedge \mathcal{T}_s$ (C_{3v}) points should be elliptic. The Morse conditions are satisfied if, out of the two three-point orbits with stabilizer $D_{2d} \times \mathcal{T}$, one contains elliptic points and the other hyperbolic points. The freedom

Table 19

Betti numbers b_k and Euler–Poincaré characteristics Σ for the spaces $\mathbb{C}P^2$ and $\mathbb{C}P^1 \sim \mathbb{S}^2$ (top). Number and type of stationary points of the simplest Morse function on the F_2-mode space $\mathbb{C}P^2$, E-mode space $\mathbb{C}P^1$, and rotational sphere $\mathbb{S}^2$ in the presence of the symmetry group $T_d \times \mathcal{T}$. The frame indicates additional stationary points of the possible nonsimplest Morse function on $\mathbb{C}P^2$.

Space	b_0	b_1	b_2	b_3	b_4	b_5	b_6	Σ
$\mathbb{C}P^1 \sim \mathbb{S}^2$	1	0	1					2
$\mathbb{C}P^2$	1	0	1	0	1			3
$\mathbb{C}P^1 \times \mathbb{S}^2$	1	0	2	0	1			4
$\mathbb{C}P^2 \times \mathbb{S}^2$	1	0	2	0	2	0	1	6
Space	c_0	c_1	c_2	c_3	c_4	c_5	c_6	Σ
$\mathbb{C}P^1$	$3D_{2d} \times \mathcal{T}$	$3D_{2d} \times \mathcal{T}$	$2T \wedge \mathcal{T}_s$					2
$\mathbb{S}^2$	$6S_4 \wedge \mathcal{T}_s$	$12C_s \wedge \mathcal{T}_2$	$8C_3 \wedge \mathcal{T}_s$					2
$\mathbb{C}P^2$	$4C_{3v} \times \mathcal{T}$	$6C_{2v} \times \mathcal{T}$	$8C_3 \wedge \mathcal{T}_s$	$6S_4 \wedge \mathcal{T}_s$	$3D_{2d} \times \mathcal{T}$			3
$\mathbb{C}P^2$	$4C_{3v} \times \mathcal{T}$	$6C_{2v} \times \mathcal{T}$	$3D_{2d} \times \mathcal{T}$ $8C_3 \wedge \mathcal{T}_s$	$\boxed{\begin{matrix} 6C_s \wedge \mathcal{T}_2 \\ 6C_s \wedge \mathcal{T}_2 \end{matrix}}$	$6S_4 \wedge \mathcal{T}_s$			3
$\mathbb{C}P^2 \times \mathbb{S}^2$	$8C^{(3)}$	$12C_s$	$6C^{(4)}$ $6A^{(4)}$	$12A^{(2)}$	$8A^{(3)}$ $8B^{(3)}$	$12C_s$	$6B^{(4)}$	6

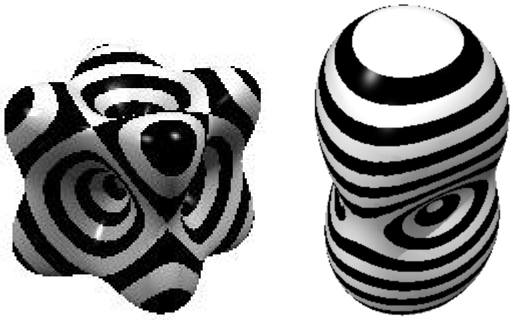


Figure 13. *Simplest purely rotational (left) and E-mode vibrational (right) Morse Hamiltonians (often called* energy surfaces*) of the* A$_4$ *molecule as functions on the phase spaces $\mathbb{S}^2$ and $\mathbb{S}^2 \sim \mathbb{C}P^1$. The bounding potential of the* A$_4$ *molecule is approximated as a sum of pairwise harmonic atom–atom interaction terms; see* [12].

of choice is limited to having two maxima and three minima or vice versa. As before, we can predict which of the two possibilities is realized in A$_4$ using the simple atom–atom vibrational potential of [12, 13]. It turns out that at fixed action n_e the two C_{3v}-symmetric RE have maximum energy; see Figure 13. Note that the same happens in the case of the E mode of the A$_3$ molecule, such as H$_3^+$ [19], whose equilibrium configuration is an isosceles triangle.

5.2.3. Morse functions on the F_2-mode phase space $\mathbb{C}P^2$. Among the fixed points of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^2$ (see Tables 10 and 17) only points with stabilizers $C_{2v} \wedge \mathcal{T}$ and $S_4 \wedge \mathcal{T}_2$ (D_2 and C_4 in the short notation) can have odd Morse indexes. Table 19 demonstrates how Morse inequalities for $\mathbb{C}P^2$ are satisfied if stationary points lie only on the critical orbits. We can interchange points of indexes 1 ↔ 3 or/and 0 ↔ 4 to obtain other possible simplest Morse functions. Table 20 shows how this simplest set of stationary points respects Morse

Table 20

Stationary points of the $(T_d \times \mathcal{T})$-invariant Morse Hamiltonians on $\mathbb{C}P^2$ (see Table 19) projected on the C_2- and C_s-invariant spheres.

Stabilizer of		Signature (index)		Number of
sphere	orbit	on $\mathbb{C}P^2$	on $\mathbb{S}^2$	points on $\mathbb{S}^2$
	Simplest Morse Hamiltonian:			
C_2	$D_{2d} \times \mathcal{T}$	$[++++]$ (0)	$[++]$ (0)	2
	$S_4 \wedge \mathcal{T}_2$	$[+++-]$ (1)	$[+-]$ (1)	2
	$C_{2v} \wedge \mathcal{T}$	$[+---]$ (3)	$[--]$ (2)	2
C_s	$D_{2d} \times \mathcal{T}$	$[++++]$ (0)	$[++]$ (0)	1
	$C_{2v} \wedge \mathcal{T}$	$[+---]$ (3)	$[+-]$ (1)	1
	$C_{3v} \times \mathcal{T}$	$[--++]$ (2)	$[--]$ (2)	2
	Nonsimplest Morse Hamiltonian:			
C_2	$D_{2d} \times \mathcal{T}$	$[+-+-]$ (2)	$[+-]$ (1)	2
	$S_4 \wedge \mathcal{T}_2$	$[++++]$ (0)	$[++]$ (0)	2
	$C_{2v} \wedge \mathcal{T}$	$[+---]$ (3)	$[--]$ (2)	2
C_s	$D_{2d} \times \mathcal{T}$	$[+-+-]$ (2)	$[+-]$ (1)	1
	$C_{2v} \wedge \mathcal{T}$	$[+---]$ (3)	$[+-]$ (1)	1
	$C_s \wedge \mathcal{T}_2$	$[+++-]$ (1)	$[++]$ (0)	2
	$C_{3v} \times \mathcal{T}$	$[--++]$ (2)	$[--]$ (2)	2

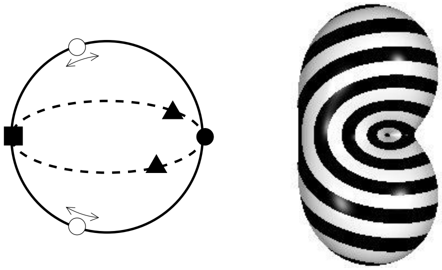


Figure 14. *Position of RE (left) and vibrational F_2-mode Hamiltonian (right) of the* A_4 *molecule restricted to the C_s-invariant sphere in the phase space $\mathbb{C}P^2$. White circles denote extra (nonfixed) RE; other markers correspond to fixed points in Figure 12 (left). The bounding potential of* A_4 *is approximated as a sum of pairwise atom–atom harmonic interaction terms* [13].

theory requirements for the C_2- or C_s-invariant spheres in Table 12 and in Figures 11 and 12 (left). It can be further verified that requirements for *all* closed invariant subspaces of $\mathbb{C}P^2$ are satisfied.

Computation with the atom–atom potential [13] (in the limit of independent vibrational and rotational motions) suggests that the simplest Morse Hamiltonian is *not* realized in real A_4 molecules. Instead we expect a Hamiltonian with 12 additional equivalent RE with stabilizer $C_s \wedge \mathcal{T}_2$, which are situated in pairs on the $(C_s \wedge \mathcal{T}_2)$-invariant main circle of the six C_s-invariant spheres, as shown in Figure 14. When the energy of the system (or the action n_f) changes, these two RE can move along the invariant circle. The set of RE of this nonsimplest system is characterized in the second to last row of Table 19 and in Table 20, bottom.

5.2.4. Morse functions on combined spaces. The Betti numbers b_k and Euler characteristics Σ for the smooth manifold P, which is a product $P' \times P''$, follow from those for factor spaces P' and P'',

$$b_k = \sum_{i+j=k} b'_i b''_j, \quad \Sigma = \sum_{k=0}^{\dim P} (-1)^k b_k,$$

where indexes i, j, and k go from 0 to $\dim P'$, $\dim P''$, and $\dim P = \dim P' + \dim P''$, respectively. In many cases we can analyze RE on P by combining the rules for P' and P''; the most interesting case turns out to be that of $\mathbb{C}P^2 \times \mathbb{S}^2$ (F_2-mode vibration and rotation).

Satisfying Morse conditions on invariant subspaces becomes increasingly important in high dimensions. Thus, even before attempting to consider whether the "minimum" set of the 12 A^{C_2}, six $(A, B, C)^{C_4}$, and eight $(A, B, C)^{C_3}$ stationary points (see Table 15—ignore indexes $_{1,2}$) satisfies all conditions for $\mathbb{C}P^2 \times \mathbb{S}^2$, we can check if this set works for the subspaces of $\mathbb{C}P^2 \times \mathbb{S}^2$. Going back to section 4.5 and Figure 12, we conclude immediately that our set is incomplete. Indeed, we should expect at least two stationary points (a maximum and a minimum) on each of the twelve C_s-invariant spheres in $\mathbb{C}P^2 \times \mathbb{S}^2$—yet *none* of the fixed points of the $T_d \times \mathcal{T}$ action lies on these spheres. Therefore, the set of stationary points on the $\mathbb{C}P^2 \times \mathbb{S}^2$ space (rotation–vibration RE) includes necessarily at least two 12-point noncritical C_s-orbits. Adding these 24 points, the simplest Morse function on $\mathbb{C}P^2 \times \mathbb{S}^2$ can be constructed; one possibility is presented in the last row of Table 19. This function corresponds to the Coriolis-dominated structure, which we will discuss on the example in section 11.

5.3. RE in the initial phase space. RE of the A$_4$ molecule in the initial phase space can be largely, and in some cases *entirely*, reconstructed using the qualitative information on the symmetry group action on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ and stationary points of the reduced Hamiltonian. Thus all purely rotational RE of a tetrahedral molecule correspond to the stationary rotation around the symmetry axis of their stabilizers. For the $S_4 \wedge \mathcal{T}$ stabilizer we take axis C_4, and for $C_s \times \mathcal{T}_2$ we take the C_2 axis orthogonal to the symmetry plane. Vibrational RE of the E- and F_2-mode systems form families of basic periodic orbits in the initial phase space parameterized by the values of integrals n_e and n_f. Rotation–vibration RE of our system are labeled by the values of integrals j, n_e, and n_f and can be reconstructed as appropriate combinations of the periodic motions of the subsystems, which correspond to the combined stationary points on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ (see Table 15) and which become 3-tori (see footnote 26) in the original phase space of the system. We will characterize these RE in more detail and illustrate the results on the concrete example of the A$_4$ molecule introduced in section 2.3 and [13].

5.3.1. RE of the E-mode system. Neglecting rotation and interaction with other vibrational modes, the E-mode system can be described by the Hamiltonian

$$H = \omega_e[H_0 + V(q)], \tag{5.1a}$$

where

$$H_0 = \frac{1}{2}\sum_{i=1}^{2}(q_{E_i}^2 + p_{E_i}^2) = \frac{1}{2}(p_{E_1}^2 + p_{E_2}^2) + V_0 \tag{5.1b}$$

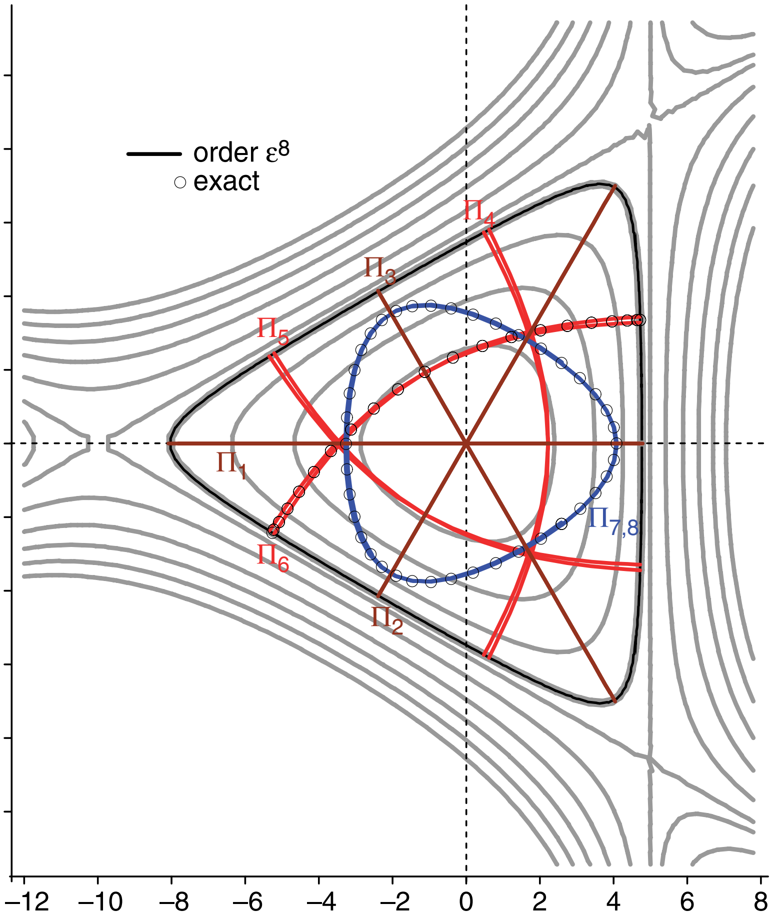


Figure 15. *Hénon–Heiles potential $V_0(q) + \varepsilon V(q)$ computed for $\varepsilon = 0.1$ and $h/h_{\text{saddle}} = 0.2$, 0.45, 0.7, 0.9 (bold contour), 1, 1.2, ... (gray). Nonlinear normal modes of the Hénon–Heiles oscillator (E-mode system) reconstructed using ϵ^8 normal form with $\varepsilon = 0.1$ and $h/h_{\text{saddle}} = 0.9$ (colored).*

is the Hamiltonian of the 1:1 harmonic oscillator, and the anharmonic part of the D_3 symmetric potential equals

$$V(q) = \varepsilon \left(\frac{1}{3} q_{E_1}^2 - q_{E_2}^2 \right) q_{E_1} + \cdots . \tag{5.1c}$$

The potential $V_0 + V$ is shown in Figure 15. We can see that to the lowest order in ε our E-mode system is equivalent to the Hénon–Heiles oscillator, which has the same symmetry $D_3 \times \mathcal{T} \sim D_{3h}$. (Note that the E-mode system of triatomic molecules with the equilateral triangle equilibrium, e.g., H_3^+ [19], also has the same symmetry and the same lowest order Hamiltonian.) In our A$_4$ example [13], $\varepsilon = -\frac{3\sqrt{3}}{4}\epsilon$ and $\omega_e = \omega$.

The RE of the E-mode system are, of course, reconstructed in the same way as the RE of the Hénon–Heiles oscillator [77, 14, 15, 16, 18, 19, 78]. We represent trajectories of this system at a given fixed energy h using their projection in the configuration space, a plane $\mathbb{R}_q^2$ with coordinates (q_{E_1}, q_{E_2}). To distinguish between trajectories with the same coordinate image, we specify their direction. The boundary of the classical motion is the h-level set of $V_0(q)+V(q)$. (We consider small amplitudes and are not interested in the unbounded motion of the Hénon–Heiles system at large energies.)

The symmetry group $T_d \times \mathcal{T}$ acts on $\mathbb{R}^2_q$ (see section 4.2) like the planar point group D_3. Operations in this group $\{1, 2C_3, 3C_2\}$ act naturally on the RE projections in the $\mathbb{R}^2_q$ space. The time reversal $\mathcal{T}$ acts trivially on the coordinate space $\mathbb{R}^2_q$ while changing signs of the momenta p and thus reversing the flow of the dynamical system. It follows that $\mathcal{T}$ changes the direction of the periodic trajectories and of their image in $\mathbb{R}^2_q$. All we should do in order to reconstruct qualitatively the projection of the RE in $\mathbb{R}^2_q$ is to suggest two curves with stabilizer $T \wedge \mathcal{T}_s$ and two groups of three curves with stabilizer $D_{2d} \times \mathcal{T}$ (see critical orbits in Table 8). In [77, 14, 15, 16] these RE are called $\Pi_{7,8}$, $\Pi_{3,4,5}$, and $\Pi_{6,7,8}$, respectively.

Since the group $(T \wedge \mathcal{T}_s)/D_2 = \{1, 2C_3, 2(C_2\mathcal{T})\}$ does not include the time reversal $\mathcal{T}$ itself, the two periodic trajectories $\Pi_{7,8}$ are mapped into each other by $\mathcal{T}$ and share the same image in $\mathbb{R}^2_q$. The image is a closed C_3-invariant loop shaped as a smoothed equilateral triangle (Figure 15, right). It is easy to check that any of the three reflections C_2 also map $\Pi_7 \leftrightarrow \Pi_8$, while the operations $C_2\mathcal{T}$ leave them invariant. The trajectories $\Pi_{3,4,5}$ and $\Pi_{6,7,8}$ project on lines (degenerate loops) in $\mathbb{R}^2_q$ because their stabilizer $(D_{2d} \times \mathcal{T})/D_2 = C_2 \times \mathcal{T}$ includes time reversal $\mathcal{T}$. Such lines should necessarily begin and end on the boundary of the motion, where the trajectory has a turning point and approaches the boundary at a right angle. This leaves two possibilities (Figure 15, right): three straight lines on the three C_2 axes and three curved lines, each intersecting one of the C_2 axes at a right angle.

5.3.2. RE of the F_2-mode system.

Neglecting rotation and interaction with other vibrational modes, the F_2-mode system can be described by the Hamiltonian

$$H = \omega_f \left[H_0 + \varepsilon V_1(q) + \varepsilon^2 H_2(q,p) + \cdots \right],$$

where

$$H_0 = \frac{1}{2}\sum_{i=1}^{3}(q_i^2 + p_i^2) = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) + V_0$$

is the Hamiltonian of the 1:1:1 harmonic oscillator and

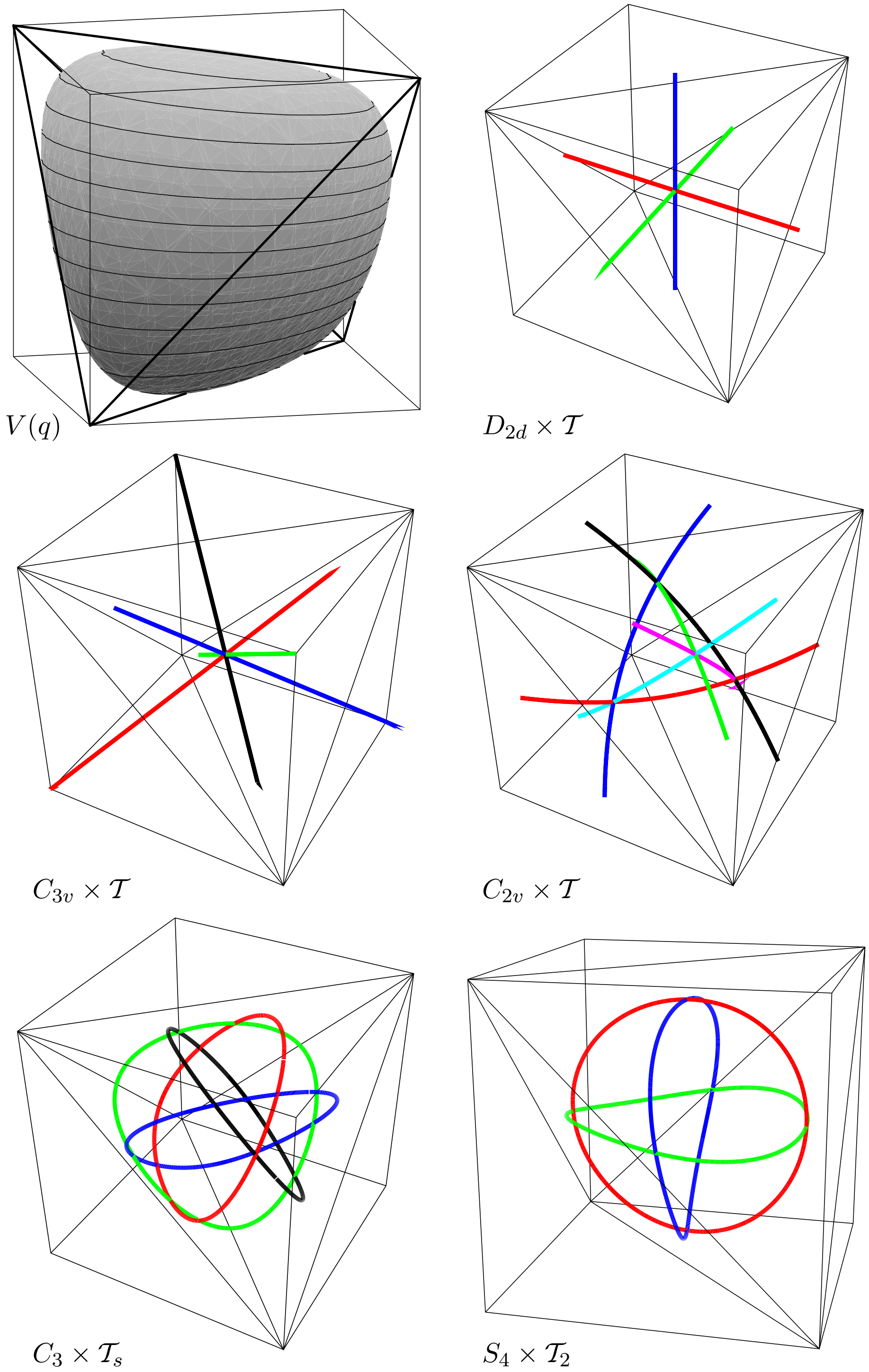
$$V_1(q) = q_1 q_2 q_3$$

is the lowest order anharmonic part of the T_d symmetric potential. The potential $V_0 + \varepsilon V_1$ illustrated in Figure 16 appears as a direct three-dimensional analogue of the two-dimensional Hénon–Heiles potential in (5.1c). However, the (small) fourth degree term

$$V_2(q) = q_1^4 + q_2^4 + q_3^4$$

should also be included for the more general description of the reduced system [98]. (Note that any T_d symmetric potential can be written as a polynomial in V_0, V_1, and V_2.) In the concrete potential of the A$_4$ molecule [13], we have

$$V(q) = V_0 + \epsilon \frac{3}{2\,(2)^{1/4}} V_1 + \epsilon^2 \left(\frac{7\sqrt{2}}{64} V_2 - \frac{5\sqrt{2}}{16} V_0^2 \right),$$



$V(q)$

$D_{2d} \times \mathcal{T}$

$C_{3v} \times \mathcal{T}$

$C_{2v} \times \mathcal{T}$

$C_3 \times \mathcal{T}_s$

$S_4 \times \mathcal{T}_2$

Figure 16. *Qualitative representation of the equipotential surface of the* F_2*-mode system (top left). Nonlinear normal modes (RE) of the three-dimensional analogue of the Hénon–Heiles oscillator (*F_2*-mode system) reconstructed for* $\epsilon = 1$ *and energy* $H_0 = 0.118$.

while $\omega_f = \sqrt{2}\omega_e$. The molecular F_2-mode Hamiltonian H_2 also contains the kinematic term $\epsilon^2 \frac{1}{8}[\mathbf{p} \times \mathbf{q}]^2$ related to the angular momentum induced by the F_2-mode vibrations.

The RE of the F_2-mode system in Figure 16 (cf. Figure 8 in [18] and [98]) can be reconstructed qualitatively using the method in the previous section. We project trajectories of this system in the configuration space $\mathbb{R}^3_q$, where they stay inside a tetrahedral cavity bounded by the h-level of $V(q)$; see Figure 16, top left. We classify curves in this cavity by their stabilizers. RE with stabilizers $D_{2d} \times \mathcal{T}$, $C_{3v} \times \mathcal{T}$, and $C_{2v} \times \mathcal{T}$ (D_4, D_3, and D_2 in the shorthand notation of section 4.3) are $\mathcal{T}$-invariant. They project to lines in $\mathbb{R}^3_q$, which are passed in both directions. Like $\Pi_{1,2,3}$ of the Hénon–Heiles system, the $D_{2d} \times \mathcal{T}$ and $C_{3v} \times \mathcal{T}$ RE lie on the corresponding symmetry axes C_4 and C_3. The $C_{2v} \times \mathcal{T}$ RE lie in the symmetry planes C_s and are slightly curved; they resemble, therefore, $\Pi_{4,5,6}$. RE with stabilizers $S_4 \times \mathcal{T}_s$ and $C_3 \wedge \mathcal{T}_s$ are similar to the "circular" RE $\Pi_{7,8}$. They project on closed directed curves in $\mathbb{R}^3_q$. In the crudest approximation, these RE can be represented as circles lying in the plane orthogonal to the respective axes C_4 and C_3. The $S_4 \times \mathcal{T}_s$ RE develops a characteristic "bow tie" twist (see Figure 16), which brings the circular symmetry down precisely to $S_4 \times \mathcal{T}_s$. The $C_3 \wedge \mathcal{T}_s$ RE has a triangular shape similar to that of $\Pi_{7,8}$ and bends slightly out of plane like the trim on a skullcap. We can further observe that the energy–action characteristics of the F_2-mode RE (see Table 19 and sections 5.1.4 and 5.2.3) also shows a certain similarity to the E-mode system: At given fixed action n_f, the energy of "circular" RE tends to be higher than that of "linear" RE.

5.4. Quantum predictions.

Quantum manifestations of RE are very familiar to physicists working on highly excited rotating molecules [31, 32, 33, 34, 18]. Within our more general context we should consider these manifestations for reduced phase spaces of dimension greater than two and products of two (or more) reduced phase spaces with two (or more) dynamical integrals of motion. The latter are quantized, and the corresponding quantum numbers label polyads or multiplets of quantum levels whose internal structure (at given fixed values of integrals) is analyzed using RE. In our system we have three dynamical integrals j, n_e, and n_f, and three corresponding quantum numbers J, N_e, and N_f, which all take integer values.

5.4.1. Systems with one dynamical integral.

The most well-known quantum "signature" of classical RE is the presence of quantum states localized predominantly near one particular stable RE (a basic stable periodic orbit). In the simplest situation, all nodes of the quantum wavefunction lie along the periodic orbit, and the number of nodes N (up to Maslov's correction μ negligible in the classical limit of large N) equals $(2\pi)^{-1}$ times the action integral taken along the orbit that in turn equals the value of the dynamical integral n for the particular RE and energy h,

$$N + \mu = n(h) = \frac{1}{2\pi} \oint_{H(p,q)=h} p\, dq,$$

where $\mu = K/2$ for a K-dimensional harmonic oscillator. The energy of such a state is as close to h, i.e., to the classical maximum or minimum energy, as possible. With excitation of oscillations about the RE growing, the nodal pattern becomes less trivial and localization disappears eventually.

Stable RE manifest themselves clearly in the structure of the energy levels. The energy level structure largely depends on the dimension of the reduced phase space P and the symmetry present. The reduced system near a stable RE on P can be represented as a nonlinear oscillator of dimension $\frac{1}{2}\dim P$. If the area of classical stability in the phase space is sufficiently large (compared to $\hbar$) we can even observe a "family" of states. If $\frac{1}{2}\dim P > 1$, the harmonic oscillator frequencies can be (partially) degenerate due to the local symmetry of the RE. In the case of k equivalent RE, our reduced system is represented locally as a k-well oscillator. The depth of the wells (or the height of the barrier) is determined by the stability of the RE. We observe k-level quasi-degenerate quantum states or *clusters*. The cluster, which is closest in energy to the classical RE limit, has the smallest splitting.

We should well distinguish the quasi-degeneracy of quantum states caused by the degeneracy of the local oscillator system and by the presence of several equivalent (by symmetry) stable RE, respectively. We also recall that the presence of the symmetry group with multi-dimensional (degenerate) irreducible representations can further complicate the analysis of the energy level patterns because quantum states with wavefunctions transforming according to rows of the same irreducible representation are strictly degenerate.

5.4.2. Examples of simple cluster structures. The most well-known molecular example of the correspondence between quantum energy levels and classical RE is the structure of individual (isolated) rotational multiplets (section 1.2). In this case $P \sim \mathbb{S}^2$, $\frac{1}{2}\dim P = 1$. We observe simple regular sequences of rotational clusters. Near the limiting RE energy, the system of almost equidistant sequences resembles a k-well one-dimensional harmonic oscillator; the energy separation between the RE and the closest (first) cluster is approximately half the distance between the clusters, i.e., half-quantum.

The $2J + 1$ multiplet of the ground vibrational state of a spherical top molecule [31, 32, 33, 34, 79] has six-fold and eight-fold clusters corresponding to stable RE (stable stationary axes of rotation) with stabilizers C_4 and C_3, respectively. The energy region near the unstable RE with stabilizer C_s separates the two cluster systems. In the case of A_4 (section 5.2) the six-fold clusters lie at the bottom energies.

Asymmetric top molecules, such as H_2O, have three paired RE. The RE in each pair correspond to classical rotation about one of the principal inertia axes in two different directions and are related by time reversal. Four RE (in two pairs) are stable and rotational levels form respective two-fold clusters (doublets).

A similar cluster structure is known for vibrational systems with $P \sim \mathbb{S}^2$, such as the E-mode system in H_3^+ [19], and in A_4 (section 5.2 and 5.3.1). Vibrational polyads of these systems are labeled by quantum number $N_e = 0, 1, \ldots$ and contain $N_e + 1$ levels (to complete the rotational analogy use $J_e = \frac{1}{2}N_e$); two-fold and three-fold clusters lie near the top and bottom polyad energies, respectively. These E-mode clusters are formed when vibrational excitation is high enough to have at least five quantum states in the polyad ($N_e > 4$). Other vibrational systems with reduced phase space $\mathbb{S}^2$ include a number of triatomic molecules with nearly 1:1 resonant stretching vibrations, notably H_2O and O_3 [22, 23, 24, 25, 26, 27]. The so-called *local modes* of these molecules are nothing else but a pair of stable equivalent RE, which bifurcates (very early) from the initial "normal mode" RE as the polyad quantum number n rises. The corresponding "local mode states" form doublets; they are commonly associated with vibrations localized on the particular atom–atom bond.

5.4.3. Quantum F_2-mode system.

So far in this section, we have summarized the fundamentals of quantum interpretation of RE that are largely known. New aspects begin here. The reduced phase space $P \sim \mathbb{C}P^2$ of the F_2-mode system is a compact space of real dimension four, and $\frac{1}{2}\dim P = 2$. This means that we have a finite number of quantum states in each *polyad* with quantum number $N_f = 0, 1, \ldots$ and that the number of states is given by a polynomial in N_f of degree 2. More precisely, the polyads of the 1:1:1 oscillator have $\frac{1}{2}(N_f + 1)(N_f + 2)$ states.

Predicting and understanding the internal structure of the F_2-mode polyads begins with the RE analysis. Taking into account Morse theory requirements for $\mathbb{C}P^2$ (see section 5.1.4 and Table 17) and its C_s- and C_2-invariant symplectic subspaces $\mathbb{S}^2$ (section 5.2.3) we can suggest stability of the set of RE in the second to last row of Table 19. One possibility is given below.

RE	Stabilizer	Signature	Stability		Type
$6C_4$	$S_4 \wedge \mathcal{T}_2$	$[----]$	ii		$B, C^{(4)}$
$12C_s$	$C_s \wedge \mathcal{T}_2$	$[---+]$	ir		Not fixed
$8C_3$	$C_3 \wedge \mathcal{T}_s$	$[++--]$	c		$B, C^{(3)}$
$3D_4$	$D_{2d} \wedge \mathcal{T}$	$[+-+-]$	rr	1:1	$A^{(4)}$
$6D_2$	$C_{2v} \wedge \mathcal{T}$	$[-+++]$	ir		$A^{(2)}$
$4D_3$	$C_{3v} \wedge \mathcal{T}$	$[++++]$	ii	1:1	$A^{(3)}$

Here, as in Table 17, we use i and r to mark imaginary and real eigenvalue pairs of the local linearized Hamiltonian; c stands for four complex eigenvalues.

Using only N_f and energy H is insufficient to untangle the rich energy level spectrum of the polyad. Since the system is not integrable (there is no third global integral), all we can do is label localized states of different kinds with different sets of additional "good" quantum numbers. When the local approximation separates into the i and/or r subsystems, quantum analysis becomes straightforward. Thus, near the two stable (elliptic) RE, which are denoted ii, our system can be represented as a two-dimensional oscillator.

At the minimum polyad energy $H(A^{(3)})$ we have an oscillator with four equivalent equilibria or "wells." Near each equilibrium it is described as a two-dimensional D_3 symmetric oscillator with 1:1 resonant harmonic frequencies. In other words, we encounter a four-fold analogue of the Hénon–Heiles system. Provided that the $A^{(3)}$ RE is sufficiently stable (the wells are deep) we may expect to find a series of "small polyads" labeled by an additional "good" quantum number $\tilde{N} = 0, 1, \ldots \ll N_f$. The structure is similar to that already discussed for the E-mode system, albeit the number of levels is quadrupled. In particular, the first level with $\tilde{N} = 0$ (the lowest level in the polyad) is a four-fold cluster. At the maximum polyad energy $H(B^{(4)})$ we find a six-well two-dimensional oscillator. The wells are C_4 symmetric and have two frequencies which are, in general, incommensurate. The level system associated with the $B^{(4)}$ RE can be described using two additional local quantum numbers $\tilde{N}'$ and $\tilde{N}''$; the first level (the highest level in the polyad) is a six-fold cluster.

5.4.4. Combined systems with several dynamical integrals.

Multiplets of combined systems are labeled with several quantum numbers. For example, rotation–vibration multiplets of the F_2-mode system are labeled with a pair of numbers (J, N_f) and contain $\frac{1}{2}(N_f+1)(N_f+2)(2J+1)$ states. The structure of such multiplets can be analyzed using our results for the individual subsystems, the principles of combining rotational and vibrational RE, and, of course, the set of critical orbits of the $T_d \times \mathcal{T}$ action given in Table 15. The new idea here is that we can continue to distinguish between the two kinds of motion, rotation and vibration, while both of them are treated classically.

In typical molecules, vibrational and rotational quanta differ by a magnitude, and the common experimental situation is that $J \gg N_f$. In this limit, it is often possible to separate the whole rotational–vibrational polyad into bands, branches, or, in the terminology of [80],[33] *vibrational components* and consider the latter for different J at fixed N_f (or/and N_e). How do RE reflect this band structure? The answer is simple: points in Table 15 with the *same* rotational coordinates (j_1, j_2, j_3) give different classical limits within the same component. Thus the F_2-mode system has three kinds of bands A, B, and C.

We recall that F_2-mode vibrations induce angular momentum $\boldsymbol{\pi}$. Rotational multiplets of the F_2-mode polyads are split into branches due to the Coriolis coupling of $\mathbf{J}$ and $\boldsymbol{\pi}$ and are labeled with the additional "good" quantum number R of the angular momentum $\mathbf{J} + \boldsymbol{\pi}$ [81]. The $N_f = 1$ fundamental state has $\boldsymbol{\pi} = 1$. This state splits into three branches with $R = J - 1,\ J,\ J + 1$, which diverge linearly as J increases. The "circular" RE of type B, C (see section 5.3.2) have maximal angular momentum $\boldsymbol{\pi}$ and are the classical limit for the $R = J \pm 1$ branches; the A-type RE have zero momentum and give the limit of the $R = 0$ branch. Provided that we add the two extra nonfixed RE of symmetry $C_s \wedge \mathcal{T}_2$ (see section 5.2.4), each classical limit branch has three types of RE with shorthand labels C_4, C_2, and C_3; the internal structure of branches can be analyzed like that of an isolated rotational state in section 5.2.1.

Similar analysis for the E-mode system shows that it has two types of branches A_1 and A_2 (see Table 15—ignore the F_2 part (z_1, z_2, z_3) and use time reversal where necessary). In particular, the $N_e = 1$ state has two branches. The splitting between them is determined by higher order rotation–vibration interactions.

The number of vibrational states and, correspondingly, the number of quantum branches, increases with vibrational excitation. The number of critical orbits and of corresponding rotational–vibrational RE remains the same. Quantum branches that lie at "intermediate energies" far from the limit given by the RE can be considered in the same way as quantum states at intermediate energies of purely vibrational polyads (sections 5.4.2 and 5.4.3), i.e., as states with more complex vibrational localization. The RE analysis of the rovibrational structure is simpler for low vibrational polyads.

The number of quantum states in each band and possible intersections of bands (vibrational components) which can change this number is the subject of further qualitative study of bands. This study is beyond the scope of the present basic RE analysis. We mention only that each band can be assigned a topological index (Chern index) [82], which gives the difference between the number of states $2J + 1$ of an isolated rotational multiplet and the number of states in the band. The sum of these indexes over all components of the polyad equals zero. Thus the number of states in the Coriolis branches of the F_2-mode fundamental state equals $2R + 1$ and the indexes are ±2 and 0. In the E-mode polyads the indexes can equal only 2 or 4 modulo 6 [80].

[33] Note that the index introduced in this work equals one-half of the Chern index introduced in [82].

6. Dynamical invariants of the reduced system.

In the previous sections we analyzed the action of the symmetry group on the reduced phase space of our system and predicted its RE entirely on the basis of this analysis. We defined RE explicitly (Table 15) in terms of coordinates $(z, \bar{z})$ of the initial system (2.1). Any given reduced Hamiltonian H_{eff} can be expressed in terms of $(z, \bar{z})$ and the energy–action characteristics of fixed RE can be computed. Stability of RE can be determined using local expansions of H_{eff}. Those RE whose position on the reduced phase space changes (as a function of energy or parameters) are found as conditional extrema of $H_{\text{eff}}(z, \bar{z})$ on the reduced phase space.

The use of initial coordinates $(z, \bar{z})$ has, however, obvious limitations. These coordinates are not well suited to studying dynamics on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. The more appropriate way to analyze the reduced system is in terms of dynamically invariant functions, which can be constructed of $(z, \bar{z})$ [44]. In the following sections we show how invariant polynomials in $(z, \bar{z})$ can be used to describe the reduced system. We will use invariants to (i) express the reduced Hamiltonian H_{eff} most compactly and unambiguously, (ii) define nonlinear coordinates on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, (iii) describe the action of the symmetry group $T_d \times \mathcal{T}$ on this space, (iv) describe the dynamics of the reduced system, and (v) characterize RE in terms of both their position on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ and stability.

We consider appropriate dynamical symmetry and its reduction in order to introduce dynamical invariants of degree 2 in $(z, \bar{z})$ (section 6.2 and Table 21), which generate the ring of all invariant polynomials. All terms in H_{eff} can be expressed as various powers of these generators. In section 6.3 we describe the structure of the ring of invariant polynomials using the Molien generating function, and later in section 6.4 we define an integrity basis in order to represent uniquely each invariant polynomial in this ring. In particular, all remaining dependence on the A_1 variables is expressed as a power series in the 1-oscillator action n_a. This happens because the A_1 vibration does not change the geometry of the molecule.

6.1. Reduction of the initial rovibrational system and normal form H_{eff}.

The zero order Hamiltonian of our system is a sum of three harmonic oscillators,

$$H_0 = \omega_{A_1} n_a + \omega_E n_e + \omega_{F_2} n_f + 0j, \tag{6.1a}$$

where n_a, n_e, n_f, and j represent oscillators with degeneracy 1, 2, 3, and 2, respectively. Explicit definition in terms of initial symplectic variables $(z, \bar{z})$ is given in Table 21. The first three oscillators describe the A_1, E, and F_2 vibrational modes, respectively. The reduced rotational subsystem is lifted to an auxiliary degenerate two-oscillator system with dynamical variables $(z_6, z_7, \bar{z}_6, \bar{z}_7)$, which is more convenient in computations.

The complete initial rotation–vibration Hamiltonian H is a power series in dynamical variables $(z, \bar{z})$,

$$H = H_0 + \epsilon H_1 + \epsilon^2 H_2 + \cdots , \tag{6.1b}$$

where ϵ is a smallness parameter, and different perturbation terms are characterized below.

Table 21

Generators of the dynamical symmetry group action describing dynamics on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$.

	Definition		Definition
n_a	$\frac{1}{2}(z_a\bar{z}_a)$	n_f	$\frac{1}{2}(z_1\bar{z}_1 + z_2\bar{z}_2 + z_3\bar{z}_3)$
		x_3	$\frac{1}{2}(z_1\bar{z}_1 - z_2\bar{z}_2)$
j	$\frac{1}{4}(z_6\bar{z}_6 + z_7\bar{z}_7)$	n_3	$\frac{1}{2}(z_3\bar{z}_3)$
j_2	$\frac{1}{4}(z_6\bar{z}_7 + z_7\bar{z}_6)$	s_1	$\frac{1}{2}(z_2\bar{z}_3 + z_3\bar{z}_2)$
j_3	$\frac{1}{4}(z_6\bar{z}_7 - z_7\bar{z}_6)i$	t_1	$\frac{1}{2}(z_2\bar{z}_3 - z_3\bar{z}_2)i$
j_1	$\frac{1}{4}(z_6\bar{z}_6 - z_7\bar{z}_7)$	s_2	$\frac{1}{2}(z_1\bar{z}_3 + z_3\bar{z}_1)$
n_e	$\frac{1}{2}(z_4\bar{z}_4 + z_5\bar{z}_5)$	t_2	$\frac{1}{2}(z_3\bar{z}_1 - z_1\bar{z}_3)i$
v_1	$-\frac{1}{4}(z_4\bar{z}_5 - z_5\bar{z}_4)i$	s_3	$\frac{1}{2}(z_1\bar{z}_2 + z_2\bar{z}_1)$
v_2	$\frac{1}{4}(z_5\bar{z}_5 - z_4\bar{z}_4)$	t_3	$\frac{1}{2}(z_1\bar{z}_2 - z_2\bar{z}_1)i$
v_3	$\frac{1}{4}(z_4\bar{z}_5 + z_5\bar{z}_4)$		

Order	Degree	Type of the term
ϵ	z^3	Cubic anharmonic terms
ϵ^2	z^4	Quartic anharmonic terms
	$z^2 j$	Coriolis interaction
	j^2	“Rigid rotor” rotation

When the frequencies ω_{A_1}, ω_E, and ω_{F_2} are incommensurate, n_a, n_e, and n_f can be regarded as three approximate integrals of motion with values N_a (see footnote 26), N_e, and N_f, respectively. The fourth integral j with the value J is the amplitude of the total angular momentum, which is strictly conserved.

We can now reduce (normalize) perturbations H_1, H_2, etc. in (6.1b) by removing all terms which do not Poisson commute with integrals n_f, n_e, and j. (Note that a priori $\{H, j\} = 0$ since j is a strict integral.) In other words, we reduce the action of the dynamical 4-torus symmetry group $\mathbb{T}^4$ (see footnote 26) on the initial 16-dimensional phase space with coordinates $(z, \bar{z})$. This group is defined by the flow of four Hamiltonian vector fields X_{n_f}, X_{n_e}, X_{n_a}, and X_j. The normalized Hamiltonian H_{eff}, also called the reduced and/or effective Hamiltonian, or simply the normal form, is invariant with regard to this flow.

6.2. Invariant polynomials of the oscillator symmetry. The dynamical symmetry of each oscillator subsystem has the form

$$\varphi : \mathbb{R}^1 \times C_k \to C_k : (t, z) \to \exp(it)\, z, \tag{6.2a}$$

where dimension k can be 3 (F_2 mode), 2 (E mode and rotation), or 1 (A_1 mode). The conjugate vector $\bar{z}$ transforms, of course, as follows:

$$(t, \bar{z}) \to \exp(-it)\, \bar{z}. \tag{6.2b}$$

It follows that all monomials in $(z, \bar{z})$ that are invariant with respect to φ are of even total degree and have the same degree in z and $\bar{z}$, e.g., $z_i\bar{z}_j$, $z_i z_j \bar{z}_m \bar{z}_l$, etc. Furthermore, all invariant polynomials can be expressed using quadratic monomials of the form $z\bar{z}$ (or similar homogeneous polynomials of degree 2), which *generate* the multiplicative ring $\mathcal{R}$ of all polynomials invariant with respect to φ. Since φ is a flow of the vector field of the linearized system with Hamiltonian

$$H_0 = \frac{1}{2} z\bar{z} = \frac{1}{2}(z_1\bar{z}_1 + z_2\bar{z}_2 + \cdots + z_k\bar{z}_k),$$

all polynomials in this ring Poisson commute with H_0, which is an integral of motion for the reduced system.

Generators are defined explicitly in Table 21. In the case of the rotational subsystem, the generators are the familiar components j_1, j_2, and j_3 of the angular momentum whose amplitude j is a constant. Since the reduced phase space of the 1:1 oscillator and that of the rotator are diffeomorphic, $\mathbb{C}P^1 \sim \mathbb{S}^2$, the generators v_1, v_2, and v_3 for the E-mode polyads can also be considered as components of an angular momentum with fixed amplitude $\frac{1}{2}n_e$ [69]. The reduced F_2-mode oscillator system is described by nine linearly independent generators. The integral of motion n_f, and polynomials x_3 and n_3, are combinations of the actions of the individual oscillators,

$$n_k = \frac{1}{2} z_k\bar{z}_k, \quad k = 1, 2, 3.$$

Invariants s and t can be considered as inner and exterior products of 2-vectors,

$$s_\alpha = \frac{1}{2}(z_\beta, \bar{z}_\beta)\cdot(z_\gamma, \bar{z}_\gamma), \qquad t_\alpha = \frac{i}{2}(z_\beta, \bar{z}_\beta)\wedge(z_\gamma, \bar{z}_\gamma).$$

This construction of invariants goes back to Weyl [83].

6.3. Generating function for oscillator symmetry. Once the action of the dynamical symmetry on the initial phase space C_k of the k-oscillator is defined explicitly in (6.2) we can compute the Molien generating function $g(\lambda)$, a heuristic tool [48, 49, 50, 84] suggesting certain structural characteristics of the ring of invariant polynomials in $(z, \bar{z})$. The function $g(\lambda)$ can be obtained directly from the Molien theorem [83]

$$g(\lambda) = \frac{1}{2\pi}\int_0^{2\pi} \frac{dt}{\det(1 - \lambda U_t)}, \tag{6.3}$$

where the $2k\times 2k$ matrix U_t represents the action of the dynamical symmetry in (6.2) on both $z_1, \ldots, z_k$ and $\bar{z}_1, \ldots, \bar{z}_N$, i.e., on all phase space variables used to construct invariants. We can see from (6.2) that U_t is a diagonal matrix

$$U_t = \operatorname{diag}(\underbrace{e^{it}, \ldots, e^{it}}_{k \text{ times}}, \underbrace{e^{-it}, \ldots, e^{-it}}_{k \text{ times}}), \tag{6.4}$$

and that

$$g(\lambda) = \frac{1}{2\pi}\int_0^{2\pi} \frac{dt}{(1 - \lambda e^{it})^k(1 - \lambda e^{-it})^k}. \tag{6.5}$$

After changing to the complex unimodular variable

$$\theta = \exp(it), \quad dt = \frac{d\theta}{i\theta}, \tag{6.6}$$

the integral (6.5) becomes a Cauchy integral

$$g(\lambda) = \frac{1}{2\pi i} \oint_{|\theta|=1} \frac{\theta^{k-1} d\theta}{(1-\lambda\theta)^k (\theta-\lambda)^k}. \tag{6.7}$$

Here we note that the formal real variable λ is used in Taylor series expansions of $g(\lambda)$ and the value of λ can be assumed arbitrarily small. In particular, we can have $|\lambda^{-1}| > 1$. Since our integral has a single pole $\theta = \lambda$ of order $k \geq 1$ within the unit circle $|\theta| = 1$, the Cauchy integral formula yields

$$g(\lambda) = \frac{1}{(k-1)!} \frac{\partial^{k-1}}{\partial\theta^{k-1}} \frac{\theta^{k-1}}{(1-\lambda\theta)^k} \bigg|_{\theta=\lambda}, \tag{6.8}$$

and in particular,[34]

$$g_{C_1/S_1}(\lambda) = 1/(1-\lambda^2), \tag{6.9a}$$

$$g_{C_2/S_1}(\lambda) = (1+\lambda^2)/(1-\lambda^2)^3, \tag{6.9b}$$

$$g_{C_3/S_1}(\lambda) = (1+4\lambda^2+\lambda^4)/(1-\lambda^2)^5, \tag{6.9c}$$

$$g_{C_k/S_1}(\lambda) = \sum_{s=0}^{k-1} \binom{k-1}{s}^2 \lambda^{2s} \bigg/ (1-\lambda^2)^{2k-1}. \tag{6.9d}$$

Here the formal variable λ represents any of the variables z and $\bar{z}$. Since all invariants are of even degree in z and $\bar{z}$, the degree in λ is also even. We can, therefore, change to variable

$$\mu = \lambda^2,$$

which represents generators in Table 21. We can also omit one factor $(1-\lambda^2)$ in the denominator of (6.9) that represents the principal oscillator invariant. Then

$$g_{\mathbb{C}P^1}(\mu) = (1+\mu)/(1-\mu)^2, \tag{6.10a}$$

$$g_{\mathbb{C}P^2}(\mu) = (1+4\mu+\mu^2)/(1-\mu)^4. \tag{6.10b}$$

6.4. Integrity basis. All functions invariant with respect to the dynamical symmetry (6.2a), and in particular the reduced Hamiltonian (the normal form) H_{eff} in (2.4), can be expressed in terms of generator invariants in Table 21. Coefficients c_k in the Taylor series for the corresponding Molien function $g(\lambda)$ at $\lambda = 0$ give the total number of linearly independent invariant polynomials of degree k. Even though the generators themselves are linearly independent, there are algebraic relations between them and the representation of c_k invariants of degree k in terms of such generators is not unique.

For example, the components of the angular momentum obey the relation

$$j_1^2 + j_2^2 + j_3^2 = j^2 = \text{const.}$$

[34] Alternative derivation of generating functions (6.9) was given in [84].

Due to this relation, the ring of all invariant polynomials generated multiplicatively by (j_1, j_2, j_3) is not free. To express any member of this ring unambiguously we can use monomials of the type $j_1^a j_2^b j_3^c$, where a and b are arbitrary nonnegative integers and c equals 0 or 1. In other words, the ring generated by (j_1, j_2, j_3) has the structure [48, 49, 50]

$$\mathcal{R}(j_1, j_2) \bullet \{1, j_3\},$$

where $\mathcal{R}$ is a polynomial ring generated freely by j_1 and j_2. This structure is described by the Molien generating function

$$g_j = (1 + \mu_j)/(1 - \mu_j)^2, \tag{6.11}$$

where the formal variable μ_j represents any of (j_1, j_2, j_3), the two denominator factors $(1-\mu_j)$ suggest two main (or *principal*) invariants of degree 1 in (j_1, j_2, j_3), while numerator terms 1 and μ_j suggest *auxiliary* invariants of degrees 0 and 1. Such decomposition of generators into principal and auxiliary is called *integrity basis*.[35] Our example shows that the choice of such basis is not unique. Thus we can equally use (j_2, j_3) and j_1. Similarly, all E-mode invariant polynomials constitute the ring

$$\mathcal{R}(v_2, v_3) \bullet \{1, v_1\}$$

described by the generating function

$$g_e = (1 + \mu_e)/(1 - \mu_e)^2. \tag{6.12}$$

Note that v_1 changes sign under time reversal $\mathcal{T}$, while v_2 and v_3 are $\mathcal{T}$-invariant. Choosing v_1 as an auxiliary (numerator) invariant is convenient for further symmetrization with respect to $\mathcal{T}$.

The choice of the integrity basis is more difficult in the case of the 1:1:1 oscillator system (F_2 mode) with the reduced space $\mathbb{C}P^2$.[36] There are nine quadratic relations ("sygyzies" of the first order) among the generators,

$$t_1^2 + s_1^2 - 4n_3n_2 = 0, \quad t_1t_2 - s_1s_2 + 2s_3n_3 = 0, \quad s_2t_3 + s_3t_2 + 2n_1t_1 = 0, \tag{6.13a}$$

$$t_2^2 + s_2^2 - 4n_3n_1 = 0, \quad t_1t_3 - s_1s_3 + 2s_2n_2 = 0, \quad s_1t_3 + s_3t_1 + 2n_2t_2 = 0, \tag{6.13b}$$

$$t_3^2 + s_3^2 - 4n_1n_2 = 0, \quad t_2t_3 - s_2s_3 + 2s_1n_1 = 0, \quad s_1t_2 + s_2t_1 + 2n_3t_3 = 0, \tag{6.13c}$$

as well as other relations of higher degree. The Molien generating function

$$g_f = (1 + 4\mu_f + \mu_f^2)/(1 - \mu_f)^4, \tag{6.14}$$

with μ_f representing any of the generators $\{x_3, n_3, s, t\}$, suggests that all four principal invariants can be chosen from $\{x_3, n_3, s, t\}$ and that there should be four auxiliary invariants of

[35] Such decomposition is known as integrity basis [83], homogeneous system of parameters [110], or Hironaka decomposition [111].

[36] The number of principal and auxiliary invariants and their degrees in $(z, \bar{z})$ can be deduced from the Molien generating function. This function, however, does not suggest the explicit construction of the generators, which may not be unique or may not be possible at all.

degree 1 and one of degree 2. The choice of four main invariants is far from arbitrary. One possible representation of the structure of this ring is

$$\mathcal{R}(x_3, s_1, s_2, s_3) \bullet \{1, n_3, n_3^2, t_1, t_2, t_3\}.$$

If we use this integrity basis, relations (6.13) should, of course, be rewritten in order to replace n_1 and n_2 as

$$n_1 = \tfrac{1}{2}(n_f - n_3 + x_3), \qquad n_2 = \tfrac{1}{2}(n_f - n_3 - x_3).$$

To obtain an unambiguous expression of the reduced rotation–vibration Hamiltonian H_{eff} defined on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, we should combine the three integrity bases introduced above. The direct multiplication of the three rings is described by the generating function

$$g = g_f g_e g_j = \frac{1 + \mu_f + 3\bar{\mu}_f + \mu_f^2}{(1-\mu_f)^4} \frac{1 + \bar{\mu}_e}{(1-\mu_e)^2} \frac{1 + \bar{\mu}_j}{(1-\bar{\mu}_j)^2},$$

where μ and $\bar{\mu}$ stand for $\mathcal{T}$-invariants and $\mathcal{T}$-covariants. We can use main invariants

$$\mathcal{R}(x_3, s_1, s_2, s_3, v_2, v_3, \underline{j_1}, \underline{j_2})$$

and auxiliary invariants

$$\begin{aligned}\{1,&\underline{v_1}, \underline{j_3}, n_3, n_3^2, \underline{t_1}, \underline{t_2}, \underline{t_3}, v_1 j_3, \underline{n_3 v_1}, \underline{n_3^2 v_1},\\ &t_1 v_1, t_2 v_1, t_3 v_1, \underline{n_3 j_3}, \underline{n_3^2 j_3}, t_1 j_3, t_2 j_3, t_3 j_3,\\ &n_3 v_1 j_3, n_3^2 v_1 j_3, \underline{t_1 v_1 j_3}, \underline{t_2 v_1 j_3}, \underline{t_3 v_1 j_3}\ \}.\end{aligned}$$

All polynomials in the above integrity basis are chosen to be either invariant or pseudoinvariant (change sign) with respect to the time reversal $\mathcal{T}$; the pseudoinvariants are underlined. This helps further symmetrization in section 7. Of course, we should multiply our ring by all integrals $\mathcal{R}(n_f, n_e, j)$. More rigorously, we should first express the normalized Hamiltonian $\mathcal{H}_{\text{nf}}$ in terms of the above integrity basis *and* $\mathcal{R}(n_f, n_e, j)$, and only then we replace n_f, n_e, j with their constant values N_f, N_e, and J, and thus obtain the *reduced* Hamiltonian H_{eff}.

7. Dynamical invariants symmetrized with respect to finite symmetries. While the integrity basis introduced above in section 6.4 serves the purpose of dynamical (oscillator) symmetry reduction, further modifications should, in principle, follow in order to take the finite symmetry of our system into account. In particular, a symmetrized basis allows us to express (2.4) using the *minimum* number of (linearly independent) terms whose coefficients can be treated by spectroscopists as free phenomenological (or "adjustable") parameters.

7.1. Symmetry properties of dynamical invariants. We first find the action of the symmetry group $T_d \times \mathcal{T}$ on the generators in Table 21. Before considering $T_d \times \mathcal{T}$, we explain the action of two basic symmetry elements, the rotation C_k and the time reversal $\mathcal{T}$.

7.1.1. Spatial axial symmetry. Consider a rotation C_φ of the Euclidean 3-space about axis 1 by angle φ, which equals $2\pi/k$ in the case of the discrete operation C_k with $k = 2, 3, \dots$. The action of C_φ on the coordinates (q_1, q_2, q_3) is defined by the familiar 3×3 orthogonal matrix

$$\left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & M(2\varphi) \end{array}\right), \quad M(\varphi) = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}. \tag{7.1}$$

Components of the angular momentum (j_1, j_2, j_3) also transform according to this matrix.

To understand the action of C_φ on the E-mode reduced space $\mathbb{C}P^1$, consider the rotation of the complex plane with coordinates (z_4, z_5) defined by matrix $M(\varphi)$ in (7.1) (recall that we rotate the (q_4, q_5) and (p_4, p_5) planes simultaneously) and show that the corresponding rotation of the 3-space with coordinates (v_1, v_2, v_3) defined in Table 21 is given by the 3×3 matrix in (7.1). It follows that for the 2-sphere $\mathbb{S}^2$, which is defined as $v_1^2 + v_2^2 + v_3^2 = \frac{1}{4}N_e^2$ and is isomorphic to $\mathbb{C}P^1$, axis v_1 is the corresponding symmetry axis of rotation. It can be equally verified that (due to the particular choice of variables (j_1, j_2, j_3) in Table 21) similar rotation of the (z_6, z_7) plane corresponds to the symmetry axis j_3.

To find how C_φ acts on the $\mathbb{C}P^2$ space, we can rotate the complex space C_3 with coordinates (z_1, z_2, z_3) using the matrix

$$\left(\begin{array}{c|c} M(\varphi) & 0 \\ \hline 0 & 1 \end{array}\right),$$

where $M(\varphi)$ is the 2×2 matrix in (7.1), and show by a direct calculation that the action of this operation on the invariants

$$\left(\frac{x_3}{\sqrt{2}}, \frac{s_3}{\sqrt{2}}\right), \quad \left(\frac{s_2}{\sqrt{2}}, \frac{s_1}{\sqrt{2}}\right), \quad \left(\frac{t_1}{\sqrt{2}}, \frac{t_2}{\sqrt{2}}\right), \quad n_3, \quad t_3,$$

is given by the matrix $\operatorname{diag}(M(2\varphi), M(\varphi), M(\varphi), 1, 1)$, i.e., that these invariants realize representations of the SO(2) group of indexes ±2, ±1, ±1, 0, and 0, respectively.

7.1.2. Time reversal symmetry $\mathcal{T}$ (or Z_2). Recalling the action of $\mathcal{T}$ on the initial vibrational variables $(z, \bar{z})$, we can see that vibrational generators v_2, v_3, s_1, s_2, s_3, x_3, and n_3 defined in Table 21 are invariants of the Z_2 action, while v_1, t_1, t_2, and t_3 are covariants,

$$(s_1, s_2, s_3) \to (s_1, s_2, s_3), \tag{7.2a}$$

$$(x_3, n_3, v_2, v_3) \to (x_3, n_3, v_2, v_3), \tag{7.2b}$$

$$(t_1, t_2, t_3, v_1) \to (-t_1, -t_2, -t_3, -v_1). \tag{7.2c}$$

Integrity basis polynomials in section 6.4 are chosen as either $\mathcal{T}$-invariant or $\mathcal{T}$-covariant (antisymmetric or antisymmetric with respect to $\mathcal{T}$). We can easily symmetrize this basis with respect to $\mathcal{T}$ by taking squares of the principal $\mathcal{T}$-covariants $\underline{j_1}$ and $\underline{j_2}$ and excluding all auxiliary covariants. The corresponding transformation of the generating function $g_f g_e g_j$ [48, 49, 50] begins with multiplying by

$$(1 + \lambda)^2/(1 + \lambda)^2 = (1 + \bar{\mu}_j)^2/(1 + \bar{\mu}_j)^2$$

(to transform the denominator) followed by expanding the numerator and sorting out all numerator terms which represent $\mathcal{T}$-invariants. The resulting generating function

$$\frac{1+\mu+19\mu^2+6\mu^3+19\mu^4+\mu^5+\mu^6}{(1-\mu)^6(1-\mu^2)^2}, \tag{7.3}$$

where μ replaces any of formal variables $\{\mu_f, \mu_e, \mu_j\}$ and $\{\bar{\mu}_f, \bar{\mu}_e, \bar{\mu}_j\}$ for $\mathcal{T}$-invariants and $\mathcal{T}$-covariants, respectively, describes polynomials on the phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, which are invariant with regard to both dynamical and time reversal symmetry. The detailed expression for the numerator of (7.3) shows the origin of the auxiliary integrity basis invariants:

$$\begin{aligned}1 &+ \mu_f + \mu_f^2 + 3\bar{\mu}_f\bar{\mu}_e + 3\bar{\mu}_e\bar{\mu}_j + 9\bar{\mu}_f\bar{\mu}_j + 3\bar{\mu}_j^2\\ &+ 3\mu_f\bar{\mu}_j^2 + 3\mu_f\bar{\mu}_e\bar{\mu}_j + 3\mu_f^2\bar{\mu}_e\bar{\mu}_j + 3\mu_f^2\bar{\mu}_j^2\\ &+ 9\bar{\mu}_f\bar{\mu}_e\bar{\mu}_j^2 + 3\bar{\mu}_f\bar{\mu}_j^3 + \bar{\mu}_e\bar{\mu}_j^3 + \mu_f\bar{\mu}_e\bar{\mu}_j^3 + \mu_f^2\bar{\mu}_e\bar{\mu}_j^3.\end{aligned}$$

7.1.3. Finite symmetry $T_d \times \mathcal{T}$. The action of $T_d \times \mathcal{T}$ on the dynamical variables in Table 21 can, in principle, be found on the basis of sections 7.1.1 and 7.1.2 if we introduce an appropriately rotated coordinate frame to study each particular axial symmetry element of the T_d group. Otherwise we can consider tensor products

$$\left[z^E \times \bar{z}^E\right]^{\Gamma} \quad \text{and} \quad \left[z^{F_2} \times \bar{z}^{F_2}\right]^{\Gamma'}$$

of vectors $z^{F_2} = (z_1, z_2, z_3)$ and $z^E = (z_4, z_5)$ that transform according to irreducible representations Γ or Γ' of T_d (and $T_d \times \mathcal{T}$) and express these products in terms of invariants in Table 21. A straightforward calculation using Clebsch–Gordan coefficients for cubic groups [85, 86, 87, 88, 89, 90, 91, 92, 93][37] gives

$$\begin{aligned}n_e &= \frac{\sqrt{2}}{2}\left[z^E \times \bar{z}^E\right]^{A_1},\\ v_1 &= -i\frac{\sqrt{2}}{4}\left[z^E \times \bar{z}^E\right]^{A_2},\\ (v_2, v_3) &= \frac{\sqrt{2}}{4}\left[z^E \times \bar{z}^E\right]^{E},\\ n_f &= \frac{\sqrt{3}}{2}\left[z^{F_2} \times \bar{z}^{F_2}\right]^{A_1},\\ \left(\frac{3n_3 - n_f}{\sqrt{3}}, x_3\right) &= -\frac{1}{\sqrt{2}}\left[z^{F_2} \times \bar{z}^{F_2}\right]^{E},\\ (t_1, t_2, t_3) &= -\frac{i}{\sqrt{2}}\left[z^{F_2} \times \bar{z}^{F_2}\right]^{F_1},\\ (s_1, s_2, s_3) &= -\frac{1}{\sqrt{2}}\left[z^{F_2} \times \bar{z}^{F_2}\right]^{F_2}.\end{aligned}$$

[37] Our parameters $h_{ff}^{\Omega(K,\Gamma)}$ correspond to $t_{ff}^{\Omega(K,\Gamma)}$ in [92] times a constant (see [13]). The values of parameters are in spectroscopic units of energy, cm^{-1}.

The transformation properties of the generators now can be obtained explicitly from the matrices in Table 4 and equations in section 7.1.2. In particular, (t_1, t_2, t_3) and (s_1, s_2, s_3) realize irreducible representations F_{1u} and F_{2g} of $T_d \times \mathcal{T}$. We can further note that the components of the 3-vectors q^{F_2}, p^{F_2}, and $z^{F_2} = q^{F_2} - ip^{F_2}$ transform according to the irreducible representation of index 1 of the 3-space rotation group SO(3). We can also show that

$$n_f, \quad (t_1, t_2, t_3), \quad \text{and} \quad \left(s_1, s_2, s_3, \frac{3n_3 - n_f}{\sqrt{3}}, x_3\right)$$

transform according to the irreducible representations of SO(3) of indexes 0, 1, and 2, respectively.

Variables (j_1, j_2, j_3) are components of the total angular momentum, which is an axial vector transforming according to the irreducible representation 1 of the SO(3) group and F_1 of the O group. We can see from (3.1c) that (j_1, j_2, j_3) realize an irreducible representation F_{1u} of $T_d \times \mathcal{T}$. (This O_h-like notation should not be confused with F_{1g}, which is the representation of the *spatial group* $O_h \subset O(3)$ realized by (j_1, j_2, j_3).)

7.2. Symmetrized integrity basis. Once the symmetry properties of the dynamical variables are established, the integrity basis in sections 6 and 7.1.2 can be symmetrized with regard to the finite group $T_d \times \mathcal{T}$ and can be used to describe the ring of all polynomials invariant with regard to $T_d \times \mathcal{T}$.

7.2.1. Symmetrized basis for the rotational subsystem. The ring $\mathcal{R}$ of polynomials in $\{j_1, j_2, j_3\}$ invariant with respect to the action of $T_d \times \mathcal{T}$ has the same structure as the ring of polynomials in $\{x, y, z\}$ invariant with respect to the action of the O_h group of transformations of $\mathbb{R}^3$. In both cases we construct an integrity basis using the components of the triply degenerate irreducible representation F_{1u} realized by $\{j_1, j_2, j_3\}$. The Molien generating function $g(A_{1g}, F_{1u}; \lambda)$ in Table 22 indicates that the ring $\mathcal{R}$ is freely generated by three invariants j^2, r_4, and r_6 of degree 2, 4, and 6, respectively. (In molecular literature these invariants have several definitions, such as Ω_4 and Ω_6 in [31, 79] and $R^{4(4,A_1)}$ and $R^{6(6,A_1)}$ in [85, 86, 87, 88, 89].) Thus, up to degree 6 in j, a purely rotational effective Hamiltonian of a tetrahedral (or octahedral) molecule has only *six* parameters corresponding to terms j^2, j^4, r_4, j^6, $j^2 r_4$, and r_6.

To express terms in the reduced rotation–vibration Hamiltonian we also need to construct Γ-covariants (i.e., polynomials that transform according to representation Γ) for all irreducible representations Γ of $T_d \times \mathcal{T}$. Corresponding Molien generating functions $g(\Gamma, F_{1u}; \lambda)$ have, of course, the same denominator as $g(A_{1g}, F_{1u}; \lambda)$ but also have a numerator num(Γ) which describes auxiliary Γ-covariants. The ring of Γ-covariants is a product of freely generated $\mathcal{R}(j^2, r_4, r_6)$ and a finite set of numerator Γ-covariants. One possible explicit choice of these covariants is suggested in Table 22. (See [94] for the discussion of integrity bases for point groups.)

Table 22

Molien functions and possible explicit definition for invariants ($\Gamma = A_{1g}$) and Γ-covariants of the action of the O_h group (and of the isomorphic group $T_d \times \mathcal{T}$) constructed from the components $\{x, y, z\}$ of the triply degenerate irreducible representation F_{1u}.

K[38]	Γ	num(Γ)[39]	Invariants and Γ-covariants[40]
0	A_{1g}	1	$x^2 + y^2 + z^2$
4			$x^4 + y^4 + z^4$
6			$x^2y^2z^2$
9	A_{1u}	λ^9	$xyz(x^2 - y^2)(y^2 - z^2)(z^2 - x^2)$
6	A_{2g}	λ^6	$(x^2 - y^2)(y^2 - z^2)(z^2 - x^2)$
3	A_{2u}	λ^3	xyz
2	E_g	$\lambda^2 + \lambda^4$	$\{\sqrt{3}(y^2 - z^2),\ y^2 + z^2 - 2x^2\}$
4			$\{\sqrt{3}(y^4 - z^4),\ y^4 + z^4 - 2x^4\}$
5	E_u	$\lambda^5 + \lambda^7$	$xyz\{y^2 + z^2 - 2x^2,\ \sqrt{3}(z^2 - y^2)\}$
7			$xyz\{y^4 + z^4 - 2x^4,\ \sqrt{3}(z^4 - y^4)\}$
4	F_{1g}	$\lambda^4 + \lambda^6 + \lambda^8$	$\{(y^2 - z^2)yz, (z^2 - x^2)zx, (x^2 - y^2)xy\}$
6			$\{(z^4 - x^4)zx, (z^4 - x^4)zx, (x^4 - y^4)xy\}$
8			$\{(x^6 - y^6)xy, (z^6 - x^6)zx, (x^6 - y^6)xy\}$
1	F_{1u}	$\lambda + \lambda^3 + \lambda^5$	$\{x,\ y,\ z\}$
3			$\{x^3,\ y^3,\ z^3\}$
5			$\{x^5,\ y^5,\ z^5\}$
2	F_{2g}	$\lambda^2 + \lambda^4 + \lambda^6$	$\{yz,\ zx,\ xy\}$
4			$\{x^2yz,\ y^2zx,\ z^2xy\}$
6			$\{x^4yz,\ y^4zx,\ z^4xy\}$
3	F_{2u}	$\lambda^3 + \lambda^5 + \lambda^7$	$\{x(y^2 - z^2),\ y(z^2 - x^2),\ z(x^2 - y^2)\}$
5			$\{x^2(y^2 - z^2),\ y^2(z^2 - x^2),\ z^2(x^2 - y^2)\}$
7			$\{x^4(y^2 - z^2),\ y^4(z^2 - x^2),\ z^4(x^2 - y^2)\}$

7.2.2. Symmetrized basis for the E-mode subsystem.

The three components $\{v_1, v_2, v_3\}$ of the induced vibrational angular momentum of the E-mode transform according to the reducible representation $A_1'' \oplus E'$ of D_{3h} (v_1 transforms according to A_1'' and $\{v_2, v_3\}$ span the doubly degenerate irreducible representation E'). The structure of the integrity basis for the invariant and Γ-covariant polynomials in $\{v_1, v_2, v_3\}$, i.e., for the functions on the vibrational E-mode phase space $\mathbb{C}P^1 \sim \mathbb{S}^2$, is described by the Molien functions $g(A_1', A_1'' \oplus E'; \mu, \lambda)$ and $g(\Gamma, A_1'' \oplus E'; \mu, \lambda)$ in Table 23.

Table 23 also suggests the explicit form of the integrity basis polynomials. One of the principal second degree invariants is, of course, the oscillator integral $\frac{1}{2}n_e = v_1^2 + v_2^2 + v_3^2$. We also note that the cubic invariant $v_2^3 - 3v_2v_3^2$ represents the three-fold symmetry and that the reduced vibrational E-mode Hamiltonian should go up to degree 6 in the initial variables

[38] Maximum index of the irreducible representation of SO(3).

[39] Molien function for Γ-covariants $g(\Gamma, F_{1u}; \lambda)$ equals

$$\frac{\text{num}(\Gamma)}{(1 - \lambda^2)(1 - \lambda^4)(1 - \lambda^6)} = \text{num}(\Gamma)\, g(A_{1g}, F_{1u}; \lambda).$$

[40] Axes $\{x, y, z\}$ correspond to symmetry axes C_4.

Table 23

Molien functions and possible explicit definition for invariants and Γ*-covariants of the action of the* D_{3h} *group (and of the isomorphic group* $(T_d \times \mathcal{T})/D_2$*) constructed from the components* $z \oplus \{x, y\}$ *of the representation* $A_1'' \oplus E'$*.*

Γ	num(Γ)[41,42]	Invariants and Γ-covariants[43,44]
A_1'	1	$x^2+y^2+z^2$, z^2, x^3-3xy^2
A_1''	μ	z
A_2'	λ^3	$3yx^2-y^3$
A_2''	$\lambda^3\mu$	$z(3yx^2-y^3)$
E'	$\lambda+\lambda^2$	$\{x,y\}$, $\{y^2-x^2, 2xy\}$
E''	$\mu(\lambda+\lambda^2)$	$\{zy, -zx\}$, $\{2xyz, z(x^2-y^2)\}$

Table 24

Molien generating functions for invariants ($\Gamma = A_{1g}$) *and* Γ*-covariants of the action of the* O_h *group (and of the isomorphic group* $T_d \times \mathcal{T}$*) constructed from the components* $\{x, y, z\}$ *of the triply degenerate irreducible representation* F_{2g}*.*

Γ[45]	num(Γ)[46,47]
A_{1g}	$1+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^9$
A_{2g}	$2\lambda^3+\lambda^4+\lambda^5+2\lambda^6$
E_g	$\lambda+2\lambda^2+\lambda^3+2\lambda^4+2\lambda^5+\lambda^6+2\lambda^7+\lambda^8$
F_{1g}	$\lambda^2+3\lambda^3+5\lambda^4+5\lambda^5+3\lambda^6+\lambda^7$
F_{2g}	$\lambda+2\lambda^2+3\lambda^3+3\lambda^4+3\lambda^5+3\lambda^6+2\lambda^7+\lambda^8$
A_{1u}	$\lambda^3+\lambda^4+2\lambda^5+\lambda^6+\lambda^9$
A_{2u}	$\lambda^2+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^7$
E_u	$\lambda^2+2\lambda^3+3\lambda^4+3\lambda^5+2\lambda^6+\lambda^7$
F_{1u}	$\lambda+\lambda^2+3\lambda^3+4\lambda^4+4\lambda^5+3\lambda^6+\lambda^7+\lambda^8$
F_{2u}	$2\lambda^2+3\lambda^3+4\lambda^4+4\lambda^5+3\lambda^6+2\lambda^7$

(q, p) in order to represent adequately the symmetry of the system.

7.2.3. Symmetrized basis for the F_2-mode subsystem. Generating functions for the invariants and covariants of the O and T_d group action on the $\mathbb{C}P^2$ are given in [84]. The generating function for the invariants has the form

$$\frac{1+2\lambda^3+3\lambda^4+3\lambda^5+2\lambda^6+\lambda^9}{(1-\lambda^2)^2(1-\lambda^3)(1-\lambda^4)}. \tag{7.4a}$$

[41] Molien function for Γ-covariants $g(\Gamma, A_1'' \oplus E'; \mu, \lambda)$ equals

$$\frac{\mathrm{num}(\Gamma)}{(1-\lambda^2)(1-\lambda^3)(1-\mu^2)} = \mathrm{num}(\Gamma)\, g(A_1', A_1'' \oplus E'; \mu, \lambda).$$

[42] Formal variables μ and λ represent z and $\{x, y\}$, respectively.

[43] In the case of the E-mode $\{z, x, y\} = \{v_1, v_2, v_3\}$.

[44] Axes z and x correspond to symmetry axes C_3 and C_2.

[45] g and u label $\mathcal{T}$ symmetric and $\mathcal{T}$ antisymmetric representations.

[46] All functions have the same denominator as in (7.4b).

[47] The formal variable λ represents $z\bar{z}$, where z is any of (z_1, z_2, z_3).

The corresponding integrity basis is further simplified due to the $\mathcal{T}$-symmetrization, which removes half of the auxiliary (numerator) invariants. The function (7.4a) becomes

$$\frac{1+\lambda^3+\lambda^4+\lambda^5+\lambda^6+\lambda^9}{(1-\lambda^2)^2(1-\lambda^3)(1-\lambda^4)}. \tag{7.4b}$$

Generating functions for covariants are given in Table 24.

Coefficients c_k of terms λ^k in the formal series expansion of the generating functions in Table 24 equal the number of linearly independent polynomials of the kind $z^k\bar{z}^k$. Thus expansion of the function (7.4b)

$$1+2\lambda^2+2\lambda^3+5\lambda^4+5\lambda^5+\cdots \tag{7.5a}$$

suggests that there are two linearly independent $(T_d\times\mathcal{T})$-invariant terms $z^3\bar{z}^3$. This does not include polynomials built with powers of the scalar n_f that can be taken into account if we divide (7.4b) by one more $(1-\lambda)$. Then the corresponding formal series

$$1+\lambda+3\lambda^2+5\lambda^3+10\lambda^4+15\lambda^5+\cdots \tag{7.5b}$$

indicates five terms of degree 3, of which three should, obviously, contain n_f. In fact there is n_f^3 and two terms of the kind $n_f z^2\bar{z}^2$.

The generators of the rings of T_d and $(T_d\times\mathcal{T})$-invariants and covariants can be constructed from the polynomials of the forms

$$\begin{bmatrix} abc \\ pqr \end{bmatrix} = z_1^a z_2^b z_3^c \bar{z}_1^p \bar{z}_2^q \bar{z}_3^r + \left\{ \begin{array}{c} \text{column} \\ \text{permutations} \end{array} \right\} \tag{7.6a}$$

and

$$\begin{pmatrix} abc \\ pqr \end{pmatrix} = \begin{bmatrix} abc \\ pqr \end{bmatrix} + \begin{bmatrix} pqr \\ abc \end{bmatrix}. \tag{7.6b}$$

In particular, $n_f = \frac{1}{2}\left[\begin{smallmatrix}100\\100\end{smallmatrix}\right]$. The Molien functions in Table 24 characterize heuristically the structure of these rings. The denominator of the function (7.4b) tells us that there are two $z^2\bar{z}^2$, one $z^3\bar{z}^3$, and one $z^4\bar{z}^4$ principal integrity basis invariants, which can enter in any degree in the expression for other invariants and covariants. The concrete choice of these four principal invariants,

$$\begin{bmatrix} 110 \\ 110 \end{bmatrix}, \quad \begin{bmatrix} 200 \\ 020 \end{bmatrix}, \quad \begin{bmatrix} 111 \\ 111 \end{bmatrix}, \quad \begin{pmatrix} 400 \\ 022 \end{pmatrix}, \tag{7.7a}$$

and five nontrivial auxiliary (numerator) invariants,

$$\begin{pmatrix} 300 \\ 120 \end{pmatrix}, \begin{pmatrix} 301 \\ 121 \end{pmatrix}, \begin{pmatrix} 410 \\ 032 \end{pmatrix}, \begin{pmatrix} 411 \\ 033 \end{pmatrix}, \begin{pmatrix} 702 \\ 144 \end{pmatrix} - \begin{pmatrix} 612 \\ 054 \end{pmatrix}, \tag{7.7b}$$

is suggested in [84].

7.2.4. Symmetrized basis for the complete system. Once we take all symmetries into consideration and *combine* the three subsystems, the integrity basis becomes very complicated. Resulting symmetrized principal polynomials and a large number of auxiliary polynomials require high powers of dynamical variables and will not be used here. Instead, we will study both the group action of $T_d \times \mathcal{T}$ and the dynamics of the reduced system in terms of simpler dynamical invariants in Table 21. In the next section we briefly describe the tensorial basis, which we use to express the effective Hamiltonian H_{eff} (normal form).

7.2.5. Tensorial bases used in molecular literature. Instead of using an integrity basis of the kind described above, spectroscopists represent their effective Hamiltonians using tensorial bases constructed by the rules of the tensorial product of the finite symmetry group of the system. For example, the F_2-mode Coriolis term is constructed as

$$\left[i \left[z^{F_2} \times \bar{z}^{F_2} \right]^{F_1} \times j^{F_1} \right]^{A_1} = -\frac{\sqrt{2}}{\sqrt{3}}(t, j).$$

(This term is invariant with regard to a larger group SO(3).) Such bases guarantee completeness but cannot exclude the possibility of linear dependence among terms of a given order. At low orders, where such dependencies are few or nonexistent, this is tolerable. Explicit construction of all linearly independent terms of a given degree using the standard coupling scheme of tensors adopted in molecular spectroscopy is often nontrivial. The difficulty increases rapidly with degree. Of course, all spectroscopic tensors can be expressed using generators in Table 21. Some of the most frequently used terms [13] are given in Tables 25 and 26.

8. Group action, fixed points, and invariant subspaces. The action of the symmetry group of our system $T_d \times \mathcal{T}$ on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ is not free. Our main interest is in the *fixed points* of this action and in the subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, which are invariant with regard to the spatial symmetry group T_d and are therefore dynamically invariant.

In section 4 we analyzed the action of $T_d \times \mathcal{T}$ using complex dynamical variables z of the initial system (see section 2.2). Below we obtain the same results using the dynamical invariants in Table 21 and their symmetry properties. These invariants serve both as dynamical variables of the reduced system and as polynomial "coordinates" on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. We use invariants in order to remove the dynamical symmetry $G_{\text{dyn}} = \mathbb{T}^4$ (see footnote 26 and section 6.1) and to avoid the ambiguity of the $(z, \bar{z})$ coordinates. (Indeed, the values of generators in Table 21 specify uniquely a G_{dyn} orbit, which corresponds to a distinct point on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$.) At the same time, we cannot label orbits of the finite group $T_d \times \mathcal{T}$ because our polynomials are not symmetrized with regard to $T_d \times \mathcal{T}$. Instead we find concrete fixed points (and invariant subspaces) for concrete stabilizer subgroups of $T_d \times \mathcal{T}$. Results are summarized in Tables 27 and 28.

8.1. Fixed points in the presence of spatial axial symmetry. We return to the general discussion of the axial symmetry in section 7.1.1. The action of any spatial rotation C_k with $k = 2, 3, \ldots, \infty$ on the $\mathbb{C}P^1$ space and on the isomorphic 2-sphere $\mathbb{S}^2$ has *two* fixed points. These points lie on the symmetry axis. Thus in the j_3 axis example,

$$j_1 = j_2 = 0, \quad j_3 = \pm J.$$

Table 25

Expression of spectroscopic tensors used in effective rotation–vibration Hamiltonians for the E and F_2 modes in terms of dynamical invariants.

H^0_{ff}	n_f
H^0_{ee}	n_e
$H^{1(1,F_1)}_{ff}$	$-\dfrac{\sqrt{2}}{\sqrt{3}}(t_3 j_3 + t_2 j_2 + t_1 j_1)$
J^2	j^2
$V^{A_1}_{eeee}$	$2\left(\dfrac{1}{4}n_e^2 - v_1^2\right)$
V^{E}_{eeee}	$\sqrt{2}\left(\dfrac{1}{4}n_e^2 + v_1^2\right)$
$V^{A_1}_{ffff}$	$\dfrac{1}{3}(s_1^2 + s_2^2 + s_3^2) + n_3^2 - \dfrac{2}{3}n_f n_3 + \dfrac{1}{3}x_3^2$
V^{E}_{ffff}	$\dfrac{\sqrt{2}}{6}\left(\dfrac{3}{2}n^2 + \dfrac{3}{2}n_3^2 + \dfrac{1}{2}x_3^2 - n_f n_3 - s_1^2 - s_2^2 - s_3^2\right)$
$V^{F_2}_{ffff}$	$\dfrac{1}{2\sqrt{3}}(n_f^2 - x_3^2 + 2n_f n_3 - 3n_3^2)$
$V^{F_1}_{efef}$	$\dfrac{1}{2\sqrt{3}}n_f n_e + \dfrac{1}{2\sqrt{3}}(3n_3 - n_f)v_2 + \dfrac{1}{2}x_3 v_3$
$V^{F_2}_{efef}$	$\dfrac{1}{2\sqrt{3}}n_f n_e - \dfrac{1}{2\sqrt{3}}(3n_3 - n_f)v_2 - \dfrac{1}{2}x_3 v_3$
$H^{2(0,A_1)}_{ff}$	$-\dfrac{4}{3}n_f j^2$
$H^{2(2,E)}_{ff}$	$\sqrt{2}(3n_3 - n_f + x_3)j_2^2 - 2\sqrt{2}\left(n_3 - \dfrac{1}{3}n_f\right)j^2$ $+\sqrt{2}(3n_3 - n_f - x_3)j_1^2$
$H^{2(2,F_2)}_{ff}$	$-\dfrac{4}{\sqrt{3}}(s_1 j_3 j_2 + s_2 j_3 j_1 + s_3 j_2 j_1)$
$H^{2(0,A_1)}_{ee}$	$-\dfrac{2\sqrt{2}}{\sqrt{3}}n_e j^2$
$H^{2(2,E)}_{ee}$	$2\sqrt{6}\,v_2\left(\dfrac{2}{3}j^2 - j_2^2 - j_1^2\right) + 2\sqrt{2}\,v_3(j_1^2 - j_2^2)$
$H^{3(3,A_2)}_{ee}$	$-16\sqrt{3}\,v_1 j_3 j_2 j_1$
J^4	j^4
$H^{4(4,A_1)}$	$16\dfrac{\sqrt{10}}{\sqrt{3}}\left(j_2^4 + j_1^4 + j_2^2 j_1^2 - j^2(j_1^2 + j_2^2) + \dfrac{1}{5}j^4\right)$

In the complex coordinates (z_6, z_7) these points can be represented as $(1,0)$ and $(0,1)$. The analysis is the same for the E-mode space.

Rotation C_k with $k > 2$ acting on the $\mathbb{C}P^2$ space has *three* fixed points. The action of this operation on the dynamical invariants is described in section 7.1.1 for the case of rotation about z_3 (take $\varphi < \pi$ because $k > 2$). In this case, we find that

$$x_3 = s_3 = s_2 = s_1 = t_1 = t_2 = 0$$

at the fixed points. Substitution into (6.13) gives

$$(1-\eta)\eta = (1-\eta)t_3 = \eta^2 - t_3^2 = 0,$$

Table 26

Relation between low degree polynomials constructed in terms of integrity basis and spectroscopic tensorial terms. Only leading terms are taken into account (i.e., classical limit commutativity of variables is assumed).

$\begin{bmatrix}100\\100\end{bmatrix}$	$2H^0_{ff}$
$\begin{bmatrix}100\\100\end{bmatrix}^2$	$4V^{A_1}_{ffff} + 4\sqrt{2}V^{E}_{ffff} + 4\sqrt{3}V^{F_2}_{ffff}$
$\begin{bmatrix}110\\110\end{bmatrix}$	$2\sqrt{3}V^{F_2}_{ffff}$
$\begin{bmatrix}200\\020\end{bmatrix}$	$8V^{A_1}_{ffff} - 4\sqrt{2}V^{E}_{ffff}$
$\begin{bmatrix}100\\100\end{bmatrix}^3$	$8\sqrt{3}\left(V^{EF_2,EF_2}_{fff,fff} + V^{F_2F_2,F_2F_2}_{fff,fff} + V^{A_1F_2,A_1F_2}_{fff,fff}\right)$
	$+24\sqrt{3}V^{EF_1,EF_1}_{fff,fff} + 8V^{F_2A_1,F_2A_1}_{fff,fff}$
$\begin{bmatrix}100\\100\end{bmatrix}\begin{bmatrix}110\\110\end{bmatrix}$	$4V^{F_2A_1,F_2A_1}_{fff,fff} + 8\sqrt{3}V^{EF_1,EF_1}_{fff,fff} + 4\sqrt{3}V^{F_2F_2,F_2F_2}_{fff,fff}$
$\begin{bmatrix}100\\100\end{bmatrix}\begin{bmatrix}200\\020\end{bmatrix}$	$8\sqrt{3}\left(2V^{A_1F_2,A_1F_2}_{fff,fff} - V^{EF_1,EF_1}_{fff,fff} - V^{EF_2,EF_2}_{fff,fff}\right)$
$\begin{bmatrix}111\\111\end{bmatrix}$	$\frac{4}{3}V^{F_2A_1,F_2A_1}_{fff,fff}$
$\begin{pmatrix}300\\120\end{pmatrix}$	$4\sqrt{3}\left(4V^{A_1F_2,A_1F_2}_{fff,fff} - V^{F_2F_2,F_2F_2}_{fff,fff} - 2V^{EF_2,EF_2}_{fff,fff}\right)$

where $n_1 = n_2 = \frac{1}{2}\eta \geq 0$ and of course $n_3 = N_f - \eta \geq 0$. This system has three solutions, all isolated fixed points on $\mathbb{C}P^2$, with $(t_3/N_f, \eta/N_f)$ equal to $(0,0)$, $(1,1)$, and $(1,-1)$, respectively. The first solution is invariant with regard to time reversal $\mathcal{T}$, while the other two constitute one $\mathcal{T}$ orbit. In the complex coordinates (z_1, z_2, z_3) these points can be represented as $(1,0,0)$ and $(0,1,\pm i)$.

In the special case of the C_2 rotation we can only assert that

$$s_2 = s_1 = t_1 = t_2 = 0.$$

This leaves two possibilities: an isolated fixed point

$$n_3 = N_f, \quad n_1 = n_2 = t_3 = s_3 = 0,$$

and a 2-sphere defined as

$$n_3 = 0, \quad t_3^2 + s_3^2 + x_3^2 = N_f^2.$$

In the original (z_1, z_2, z_3) coordinates, the former is again the point $(1,0,0)$, while the latter is the $\mathbb{C}P^1$ subspace of $\mathbb{C}P^2$, where $z_3 = 0$.

Table 27

Points in the critical orbits of the $T_d \times \mathcal{T}$ group action on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ characterized by values of dynamical invariants in Table 21. Points are listed without their time reversal ($\mathcal{T}$) companions; upper and lover signs in the ± and ∓ notation correspond to different orbits with subscript indices 1 and 2, respectively. Matrices of stabilizers S_4^z, $C_3^{[111]}$, and C_s^{xy} are given in Table 4.

Orbit	$\frac{x_3}{N_f}$	$\frac{s_1}{N_f}$	$\frac{s_2}{N_f}$	$\frac{s_3}{N_f}$	$\frac{n_3}{N_f}$	$\frac{t_1}{N_f}$	$\frac{t_2}{N_f}$	$\frac{t_3}{N_f}$	$\frac{v_1}{N_e}$	$\frac{v_2}{N_e}$	$\frac{v_3}{N_e}$	$\frac{j_1}{J}$	$\frac{j_2}{J}$	$\frac{j_3}{J}$	Stabilizer
$A_{1,2}^{(2)}$	0	0	0	−1	0	0	0	0	0	$\mp\frac{1}{2}$	0	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	0	C_s^{xy}
	0	0	0	1	0	0	0	0	0	$\mp\frac{1}{2}$	0	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	0	
	$\frac{1}{2}$	0	−1	0	$\frac{1}{2}$	0	0	0	0	$\pm\frac{1}{4}$	$\pm\frac{\sqrt{3}}{4}$	$\frac{1}{\sqrt{2}}$	0	$\frac{-1}{\sqrt{2}}$	
	$\frac{1}{2}$	0	1	0	$\frac{1}{2}$	0	0	0	0	$\pm\frac{1}{4}$	$\pm\frac{\sqrt{3}}{4}$	$\frac{1}{\sqrt{2}}$	0	$\frac{1}{\sqrt{2}}$	
	$-\frac{1}{2}$	1	0	0	$\frac{1}{2}$	0	0	0	0	$\pm\frac{1}{4}$	$\mp\frac{\sqrt{3}}{4}$	0	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	
	$-\frac{1}{2}$	−1	0	0	$\frac{1}{2}$	0	0	0	0	$\pm\frac{1}{4}$	$\mp\frac{\sqrt{3}}{4}$	0	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	
$A_{1,2}^{(3)}$	0	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0	0	0	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$C_3^{[111]}$
	0	$\frac{2}{3}$	$-\frac{2}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$	0	0	0	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$-\frac{2}{3}$	$\frac{2}{3}$	$-\frac{2}{3}$	$\frac{1}{3}$	0	0	0	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$-\frac{2}{3}$	$-\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0	0	0	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	
$B_{1,2}^{(3)}$	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$C_3^{[111]}$
	0	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	
$C_{1,2}^{(3)}$	0	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$C_3^{[111]}$
	0	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	
	0	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\mp\frac{1}{2}$	0	0	$\frac{-1}{\sqrt{3}}$	$\frac{-1}{\sqrt{3}}$	$\frac{1}{\sqrt{3}}$	
$A_{1,2}^{(4)}$	0	0	0	0	1	0	0	0	0	$\mp\frac{1}{2}$	0	0	0	1	S_4^z
	1	0	0	0	0	0	0	0	0	$\pm\frac{1}{4}$	$\mp\frac{\sqrt{3}}{4}$	1	0	0	S_4^x
	−1	0	0	0	0	0	0	0	0	$\pm\frac{1}{4}$	$\pm\frac{\sqrt{3}}{4}$	0	1	0	S_4^y
$B_{1,2}^{(4)}$	0	0	0	0	0	0	0	1	0	$\mp\frac{1}{2}$	0	0	0	1	S_4^z
	$-\frac{1}{2}$	0	0	0	$\frac{1}{2}$	1	0	0	0	$\pm\frac{1}{4}$	$\mp\frac{\sqrt{3}}{4}$	1	0	0	S_4^x
	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0	1	0	0	$\pm\frac{1}{4}$	$\pm\frac{\sqrt{3}}{4}$	0	1	0	S_4^y
$C_{1,2}^{(4)}$	0	0	0	0	0	0	0	−1	0	$\mp\frac{1}{2}$	0	0	0	1	S_4^z
	$-\frac{1}{2}$	0	0	0	$\frac{1}{2}$	−1	0	0	0	$\pm\frac{1}{4}$	$\mp\frac{\sqrt{3}}{4}$	1	0	0	S_4^x
	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	0	−1	0	0	$\pm\frac{1}{4}$	$\pm\frac{\sqrt{3}}{4}$	0	1	0	S_4^y

8.2. Fixed points of the $T_d \times \mathcal{T}$ group action.

8.2.1. Stabilizer S_4.

Orientation of the S_4^z axis in Table 4 corresponds to the one used in section 8.1 for the general case of a C_k axis; solutions for fixed points with stabilizer C_4 on the F_2-mode space $\mathbb{C}P^2$ and on the rotational space $\mathbb{S}^2$ are already given above in section 8.1. On the E-mode space $\mathbb{C}P^1 \sim \mathbb{S}^2$; the image of the S_4^z operation is a rotation about axis v_2 by

Table 28

Invariant subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ with stabilizers C_s^{xy} and C_2^z.

	Values of dynamical variables and defining equation(s)															
Stabilizer	$\frac{x_3}{N_f}$	$\frac{s_1}{N_f}$	$\frac{s_2}{N_f}$	$\frac{s_3}{N_f}$	$\frac{n_3}{N_f}$	$\frac{t_1}{N_f}$	$\frac{t_2}{N_f}$	$\frac{t_3}{N_f}$	$\frac{v_1}{N_e}$	$\frac{v_2}{N_e}$	$\frac{v_3}{N_e}$	$\frac{j_1}{J}$	$\frac{j_2}{J}$	$\frac{j_3}{J}$	Orbit size	Topology
C_2^z	0	0	0	0	1	0	0	0	ν_1	ν_2	ν_3	0	0	1	6	
	$\nu_1^2 + \nu_2^2 + \nu_3^2 = \frac{1}{4}$															S_2
C_2^z	ξ	0	0	σ	0	0	0	τ	ν_1	ν_2	ν_3	0	0	1	6	
	$\tau^2 + \sigma^2 + \xi^2 = 1, \quad \nu_1^2 + \nu_2^2 + \nu_3^2 = \frac{1}{4}$															$S_2 \times S_2$
C_s^{xy}	0	σ	σ	η	$1-\eta$	τ	$-\tau$	0	0	$\frac{1}{2}$	0	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	0	12	
	$2\tau^2 + 2\sigma^2 + (2\eta - 1)^2 = 1$															S_2
	0	σ	σ	η	$1-\eta$	τ	$-\tau$	0	0	$-\frac{1}{2}$	0	$\frac{1}{\sqrt{2}}$	$\frac{-1}{\sqrt{2}}$	0	12	

angle π. Consequently, at the two fixed points on this space, $v_1 = v_3 = 0$.

8.2.2. Stabilizer C_2. We can consider the C_2^z stabilizer using directly the results in section 8.1. There is a fixed point and an invariant 2-sphere in $\mathbb{C}P^2$ and two fixed points on the rotational sphere $\mathbb{S}^2$ (the fixed points are the same as in the case of S_4^z). Furthermore, the whole E-mode space $\mathbb{C}P^1 \sim \mathbb{S}^2$ is invariant because operations C_2^z, C_2^x, and C_2^y act trivially on this space. There is, therefore, no restriction on (v_1, v_2, v_3). On the full reduced space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ we can have a point, an invariant 2-sphere, or an $\mathbb{S}^2 \times \mathbb{S}^2$ space.

8.2.3. Stabilizer C_3. In the case of C_3, we also expect three fixed points on $\mathbb{C}P^2$ and two points on $\mathbb{C}P^1$ and $\mathbb{S}^2$ each. We first recall that (j_1, j_2, j_3) transform according to the irreducible representation F_1 of the T_d group. Considering the matrix representation of F_1 for the particular operation $C_3^{[111]}$ in Table 4, we can see immediately that a point on the rotational sphere $\mathbb{S}^2$ remains invariant (stable) with respect to this operation only if

$$j_1 = j_2 = j_3.$$

Furthermore, since

$$j_1^2 + j_2^2 + j_3^2 = j^2 = J^2,$$

the two possible solutions for the fixed points on $\mathbb{S}^2$ are

$$j_1 = j_2 = j_3 = \pm J/\sqrt{3}.$$

These two points form one $\mathcal{T}$ orbit. There are four axes C_3 (four conjugate subgroups C_{3v} of the T_d group) and there is one orbit of the action of the full group $T_d \times \mathcal{T}$ that includes all eight points on $\mathbb{S}^2$ with stabilizer C_3.

Vibrational E-mode polynomials v_1 and (v_2, v_3) are chosen so that they transform according to the irreducible representations A_2 and E of T_d (see section 7). When the E-mode reduced phase space is defined using equation

$$v_1^2 + v_2^2 + v_3^2 = \frac{1}{4} n_e^2 = \frac{1}{4} N_e^2$$

as a 2-sphere in the ambient 3-space with coordinates (v_1, v_2, v_3), the action of the C_3 operation is equivalent to the C_3 rotation about axis v_1. (This illustrates the abstract statement that the image of the T_d group in this case is a dihedral group D_3.) The two points that are invariant with regard to this operation lie on the v_1 axis (on the diametrically opposite ends),

$$v_1 = \pm\frac{1}{2}N_e, \quad v_2 = v_3 = 0.$$

Since v_1 has the symmetry A_2, these points are mapped into each other by operations S_4 and C_s of T_d and, therefore, they are equivalent and form one two-point orbit. Operation $\mathcal{T}$ also maps these points into each other.

To find the fixed points on $\mathbb{C}P^2$ with stabilizer $C_3^{[111]}$, we note that polynomials (s_1, s_2, s_3) and (t_1, t_2, t_3) transform according to the irreducible representations F_2 and F_1 of the T_d group, respectively. From matrices in Table 4 we conclude that at the fixed points on $\mathbb{C}P^2$,

$$s_1 = s_2 = s_3 = \sigma N_f, \qquad t_1 = t_2 = t_3 = \tau N_f, \tag{8.1a}$$

where σ and τ are dimensionless. We further note that at the same fixed point (with stabilizer $C_3^{[111]}$) polynomials

$$\left(\frac{3n_3 - n_f}{\sqrt{3}},\, x_3\right),$$

which transform according to the irreducible representation E, should vanish, i.e.,

$$3n_3 - n_f = x_3 = n_1 - n_2 = 0, \tag{8.1b}$$

and since $n_1 + n_2 + n_3 = N_f$ we obtain

$$n_1 = n_2 = n_3 = \frac{1}{3}N_f. \tag{8.1c}$$

Substituting conditions (8.1) into relations (6.13) produces equations

$$\left\{(1+3\sigma)\tau = 0, \quad \sigma^2 + \tau^2 = \frac{4}{9}, \quad \sigma^2 - \tau^2 = \frac{2}{3}\sigma\right\}$$

with two kinds of solutions:

$$(\tau, \sigma) = \left(0, \frac{2}{3}\right) \quad \text{and} \quad \left(\pm\frac{1}{\sqrt{3}}, -\frac{1}{3}\right).$$

Combining fixed points on $\mathbb{S}^2$, $\mathbb{C}P^1$, and $\mathbb{C}P^2$ for the *same* stabilizer, i.e., the group generated by the $C_3^{[111]}$ operation in Table 4, we obtain the fixed points on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ listed in Table 27.

8.2.4. Stabilizer C_s. In the case of C_s, solution for the fixed points on the rotational sphere $\mathbb{S}^2$ and on the E-mode space $\mathbb{C}P^1 \sim \mathbb{S}^2$ is again quite simple. Indeed, using F_1 and E matrices of the C_s^{xy} example in Table 4 we find that these fixed points are defined as

$$j_1 = -j_2, \quad j_3 = 0 \quad \text{and} \quad v_1 = 0, \quad v_3 = 0.$$

On the F-mode space $\mathbb{C}P^2$ we look for a fixed point and a C_s-invariant sphere (see section 8.1). In the particular case of the C_s^{xy}-invariant points we find that

$$t_3 = 0, \quad t_1 = -t_2 = \tau N_f, \quad s_1 = s_2 = \sigma N_f, \quad x_3 = 0.$$

Using the notation

$$n_1 = n_2 = \frac{\eta}{2} N_f \geq 0, \quad n_3 = (1-\eta) N_f \geq 0, \quad s_3 = s N_f,$$

we obtain from relations (6.13) that

$$\begin{aligned} \tau^2 + \sigma^2 &= 2(1-\eta)s = 2(1-\eta)\eta, \\ \eta^2 - s^2 &= \tau(\eta - s) = \sigma(\eta - s) = 0. \end{aligned}$$

These equations have two kinds of solutions: an isolated point with

$$\eta = 1, \quad s = -1, \quad \sigma = \tau = 0,$$

which is listed in Table 27, and a 2-sphere defined by

$$\eta = s, \quad 2\tau^2 + 2\sigma^2 + (2\eta - 1)^2 = 1.$$

8.3. Orbits of the $T_d \times \mathcal{T}$ action. To find the *orbits* of equivalent fixed points of the $T_d \times \mathcal{T}$ action we take the fixed points found in the previous section for concrete stabilizers in Table 4 and act on them by all symmetry operations of $T_d \times \mathcal{T}$. We use the symmetry properties of dynamical invariants (generators in Table 21) described in section 7.1.3. These invariants are not symmetrized with regard to $T_d \times \mathcal{T}$ and their values (which play the role of "coordinates" on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$) differ for the points in the orbit. On the contrary, the value and behavior of any $T_d \times \mathcal{T}$ invariant function, such as the reduced Hamiltonian H_{eff}, remains the same.

Table 27 presents orbits of the $T_d \times \mathcal{T}$ action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. The list of fixed points in each orbit starts with the particular point found in section 8.2. Since time reversal images of points can be easily found using (7.2), we omit them for brevity so that each orbit in Table 27 has twice the number of points listed. For example, the orbit $A_1^{(4)}$ with stabilizer $S_4 \times \mathcal{T}$ has six equivalent fixed points, which correspond to three conjugate symmetry operations S_4^z, S_4^x, and S_4^y of the T_d group. The other six-point orbit $A_2^{(4)}$ differs from $A_1^{(4)}$ in the way the F_2- and E-mode coordinates are combined.

Invariant subspaces in Table 28 also are representatives of orbits of equivalent subspaces. There are six 2-spheres with stabilizer C_2, six $\mathbb{S}^2 \times \mathbb{S}^2$ spaces with the same stabilizer, and two different orbits of twelve 2-spheres with stabilizer C_s. Explicit coordinate representations for all these spaces can be obtained in the same way as obtained for the fixed points.

Table 29

Further stratification of the invariant subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ with stabilizers C_s^{xy} and C_2^z (Table 28) due to residual group action.

Stabilizer	Equations	Topology
$C_2^z \times \mathcal{T}_2$	$v_1 = 0$	S_1
$C_2^z \times \mathcal{T}_s$	$v_2 = 0$	S_1
$C_2^z \times \mathcal{T}_2$	$v_1 = x_3 = \xi = 0$	$T_2 = S_1 \times S_1$
$C_2^z \times \mathcal{T}_s$	$v_2 = s_3 = \sigma = 0$	$T_2 = S_1 \times S_1$
$C_s^{xy} \times (\mathcal{T}_2, \mathcal{T}_s)$	$\sigma = 0$	S_1

8.4. Residual group action on invariant subspaces.

Invariant subspaces of the T_d group action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, which are characterized in Table 28, are not homogeneous spaces with regard to the $T_d \times \mathcal{T}$ action. This has, of course, important consequences for the dynamics, which we will analyze later.

First we can verify whether some of the fixed points of the $T_d \times \mathcal{T}$ group action in Table 27 lie on any of the invariant subspaces. The presence of fixed points indicates that there is some nontrivial residual action of $T_d \times \mathcal{T}$ on the subspace. Continuing the C_2^z example in Table 28, we find that two points $A_1^{(4)}$ and $A_2^{(4)}$ lie on the C_2-invariant E-mode sphere, and four points $B_1^{(4)}$, $B_2^{(4)}$, $C_1^{(4)}$, and $C_2^{(4)}$ lie on the C_2-invariant space $\mathbb{S}^2 \times \mathbb{S}^2$. In the particular case of C_2^z, the residual T_d action is equivalent to the rotation by π about axes t_3 and v_2 in the respective ambient 3-spaces.

A complete study of all residual symmetries can be easily done by selecting all operations of the $T_d \times \mathcal{T}$ group which map invariant subspaces into themselves. Such selection is, of course, greatly simplified by the fact that many invariants take definite fixed values (see Table 28). Thus when studying symmetry operations acting on the invariant space $\mathbb{S}^2 \times \mathbb{S}^2$ with stabilizer C_2^z, we consider only those operations of $T_d \times \mathcal{T}$ which leave j_3 invariant (such as rotations around axis 3) or change its sign (such as reflections C_s in the planes containing axis 3). In the latter case, we should add the $\mathcal{T}$ operation to restore the sign of j_3. Another simplifying observation is that invariants used as coordinates on the subspaces often transform according to (rows of) different irreducible representations of the T_d group. In that case the residual group can have only one-dimensional representations.

Residual group action of $T_d \times \mathcal{T}$ on the invariant subspaces of the action of the spatial group T_d on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ is characterized in Table 30. We can see that the residual action on the C_2 and C_s invariant spaces is equivalent to that of a C_{2v} and a C_h group, respectively. (As before we use point group analogies of groups which include reversing operations $\mathcal{T}$, $\mathcal{T}_s$, or $\mathcal{T}_2$.) Due to the presence of this residual action, invariant subspaces contain lower-dimensional strata in addition to just fixed points; see Tables 29 and 30.

9. Dynamics of the reduced system.

Classical equations of motion for the reduced system can, in principle, be obtained using initial dynamical variables $(z, \bar{z})$ and the Poisson bracket

$$\{z, \bar{z}\} = \{q + ip, q - ip\} = -2i$$

Table 30

Residual action of $T_d \times \mathcal{T}$ on the dynamical invariants used to represent invariant subspaces in Table 28; note that $\mathcal{T}_2$ and $\mathcal{T}_s$ stand for $C_2 \circ \mathcal{T}$ and $C_s \circ \mathcal{T}$, respectively.

Spaces with stabilizer C_2^z

Dynamical		Classes of $T_d \times \mathcal{T}$			
invariants		I, C_2	$\mathcal{T}_2$	S_4	$\mathcal{T}_s$
$F_{1u}^{(3)}$	t_3, j_3	1	1	1	1
$E_g^{(1)}$	v_2, n_3	1	1	1	1
$E_g^{(2)}$	x_3, v_3	1	1	-1	-1
$F_{2g}^{(3)}$	s_3	1	-1	-1	1
A_{2u}	v_1	1	-1	-1	1

2-sphere with stabilizer C_s^{xy}

Dynamical		Classes	
invariants		I, C_s	$\mathcal{T}_2$
$E_g^{(1)}$	v_2, n_3	1	1
F_{1u}	$\dfrac{t_1 - t_2}{\sqrt{2}}, \dfrac{j_1 - j_2}{\sqrt{2}}$	1	1
F_{2g}	$\dfrac{s_1 + s_2}{\sqrt{2}}$	1	-1

if we consider the reduced Hamiltonian H_{eff} as a function of $(z, \bar{z})$. These equations of motion should preserve all symplectic symmetries of H_{eff}, i.e., remain invariant with regard to the dynamical (oscillator) symmetry and to the action of the spatial group T_d described in section 7. We can, therefore, represent them in terms of invariants in Table 21 and thus obtain equations of motion on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. A more elegant approach is to study the Poisson algebra generated by the invariants and then obtain the same equations directly.

9.1. Poisson algebra of dynamical invariants. Invariants in Table 21 generate a multiplicative ring $\mathcal{R}$ of polynomials invariant with regard to the oscillator symmetry. The Poisson bracket of any two generators in Table 21 is itself a dynamically invariant polynomial function, which is a member of $\mathcal{R}$. We say that $\mathcal{R}$ has a Poisson structure. Invariant polynomials in Table 21 generate a Poisson algebra and can be used as dynamical variables. The integrals j, n_e, and n_f are Casimirs. To compute the structure of the algebra, we can return to the $(z, \bar{z})$ representation.

For the E-mode invariants and, of course, for the angular momentum components, we obtain the standard algebra so(3),

$$\{j_\alpha, j_\beta\} = \epsilon_{\alpha\beta\gamma}\, j_\gamma, \quad \{v_\alpha, v_\beta\} = \epsilon_{\alpha\beta\gamma}\, v_\gamma,$$

and the Euler–Poisson equations,

$$\frac{d}{dt} j_\alpha = \{H_{\text{eff}}, j_\alpha\} = \frac{\partial H_{\text{eff}}}{\partial j_\gamma} j_\beta - \frac{\partial H_{\text{eff}}}{\partial j_\beta} j_\gamma. \tag{9.1a}$$

Dynamics on the $\mathbb{C}P^2$ space is described by the system of equations for eight invariant polynomials, which can be considered as independent dynamical variables. Since all these polynomials are quadratic in $(z, \bar{z})$, their Poisson brackets are also quadratic and can be expressed as their linear combinations. Resulting Poisson algebra is characterized in Table 31. Given the structure matrix $\mathcal{M}$ in this table, equations of motion can be written as

$$\dot{\theta} = \mathcal{M}^T \nabla_\theta H_{\text{eff}}, \tag{9.1b}$$

where θ is a vector $\theta = (x_3, s_1, s_2, s_3, n_3, t_1, t_2, t_3)$ and $\mathcal{M}^T = -\mathcal{M}$.

Table 31

Poisson algebra of the invariants describing dynamics on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ *(*F_2*- and* E*-mode polyads and rotational subsystem). Here* $x_1 = n_2 - n_3$*,* $x_3 = n_1 - n_2$*,* $x_2 = n_3 - n_1$*, and* $x_1 + x_2 + x_3 = 0$*.*

	s_1	s_2	s_3	n_3	t_1	t_2	t_3
x_3	$-t_1$	$-t_2$	$2t_3$	0	s_1	s_2	$-2s_3$
s_1		$-t_3$	t_2	t_1	$2x_1$	$-s_3$	s_2
s_2			$-t_1$	$-t_2$	s_3	$2x_2$	$-s_1$
s_3				0	$-s_2$	s_1	$2x_3$
n_3					s_1	$-s_2$	0
t_1						t_3	$-t_2$
t_2							t_1

	j_2	j_3
j_1	j_3	$-j_2$
j_2		j_1

	v_2	v_3
v_1	v_3	$-v_2$
v_2		v_1

9.2. Canonical variables in the limit of linearization. In section 8.2 we found RE, or equilibria of the reduced system, as isolated fixed points (critical orbits) of the $T_d \times \mathcal{T}$ action. To study dynamics near these RE, and in particular to determine their stability, we should linearize H_{eff} near them. Linearization in terms of $(z, \bar{z})$ variables was already introduced in section 5.1.4.

Similarly to finding coordinates of RE in section 8.2, linearizing near an RE can be understood on the example of a C_k symmetric RE whose symmetry axis is oriented as in section 7.1.1. Axis S_4^z (or C_4) has such an orientation, and coordinates of the RE $A_{1,2}^{(4)}$ and $B_{1,2}^{(4)}$ on the $\mathbb{C}P^2$ space and rotational sphere $\mathbb{S}^2$ (see Table 27) define, in fact, fixed points and RE for *any* C_k^z action on these spaces with $k = 3, 4, \ldots, \infty$.

Consider first the familiar simple case of the rotational sphere $\mathbb{S}^2$ described by (j_1, j_2, j_3) or by scaled variables

$$\tilde{j}_i = \frac{j_i}{\sqrt{J}}, \quad i = 1, 2, 3. \tag{9.2}$$

At the fixed point $j_3 = J$, the only nonzero Poisson bracket is $\{j_1, j_2\} = j_3$ (Table 31), and consequently the scaled variables $(\tilde{j}_1, \tilde{j}_2)$ become the standard canonical coordinate–momentum pair in the limit of linearization near the relative equilibrium with $j_3 = J$. Near the second fixed point on $\mathbb{S}^2$ with $j_3 = -J$ (which is the time reversal image of the first point) canonical variables will be $(\tilde{j}_2, \tilde{j}_1)$; i.e., $\tilde{j}_2$ will play the role of coordinate and $\tilde{j}_1$ the role of conjugate momentum.

The E-mode space $\mathbb{C}P^1$ is isomorphic to a sphere $\mathbb{S}^2$ defined in the ambient 3-space with coordinates (v_1, v_2, v_3). We should scale these coordinates as follows:

$$\tilde{v}_i = \frac{v_i \sqrt{2}}{\sqrt{N_e}}, \quad i = 1, 2, 3. \tag{9.3}$$

The S_4^z operation acts on the E-mode sphere as rotation by π about axis v_2; the two fixed points with $v_2 = \pm N_e/2$ lie on this axis. In the limit of linearization near these points we use canonical coordinates $(\tilde{v}_1, \tilde{v}_3)$ and $(\tilde{v}_3, \tilde{v}_1)$.

On the F_2-mode space $\mathbb{C}P^2$ we proceed in a similar fashion [97]. We compute the Poisson structure in Table 31 at each relative equilibrium and then find canonical variables of the

Table 32

Standard canonical coordinates and conjugate momenta for the linearization near RE with stabilizer S_4^z.

Space	Mode[48]	$A_1^{(4)}$	$B_1^{(4)}$	$\overline{B}_1^{(4)}$
$\mathbb{C}P^2$	F_2	$\tilde{t}_1, \tilde{s}_1$	$\dfrac{\tilde{s}_2 - \tilde{t}_1}{\sqrt{2}}, \dfrac{\tilde{s}_1 - \tilde{t}_2}{\sqrt{2}}$	$\dfrac{\tilde{s}_1 + \tilde{t}_2}{\sqrt{2}}, \dfrac{\tilde{s}_2 + \tilde{t}_1}{\sqrt{2}}$
		$\tilde{s}_2, \tilde{t}_2$	$\tilde{x}_3, \tilde{s}_3$	$\tilde{s}_3, \tilde{x}_3$
$\mathbb{C}P^1$	E	$\tilde{v}_1, \tilde{v}_3$	$\tilde{v}_1, \tilde{v}_3$	$\tilde{v}_1, \tilde{v}_3$
$\mathbb{S}^2$	rot	$\tilde{\jmath}_1, \tilde{\jmath}_2$	$\tilde{\jmath}_1, \tilde{\jmath}_2$	$\tilde{\jmath}_2, \tilde{\jmath}_1$

linearization limit. After changing to scaled variables

$$\tilde{a}_i = \frac{a_i}{\sqrt{2N_f}}, \quad a = s, t, n, x, \quad i = 1, 2, 3, \tag{9.4}$$

we obtain the results in Table 32.

Another way to proceed is to find coordinates near each RE on $\mathbb{C}P^2$ without demanding that these coordinates be canonical. The set of coordinate invariants $\zeta_i^{(c)}$ $(i = 1, \ldots, 4)$ that we use is selected from among the invariants n_i, s_i, and t_i $(i = 1, 2, 3)$ such that the syzygy relations (6.13), together with the constraint $n_1 + n_2 + n_3 = N_f$, can be solved in order to express the remaining invariants $\zeta_i^{(r)}$ $(i = 1, \ldots, 5)$ in terms of $\zeta_i^{(c)}$ and N_f.

We find a set of invariants with the above property by checking that the conditions of the implicit function theorem are satisfied. Specifically, let F_i $(i = 1, \ldots, 9)$ be the left-hand sides of the syzygy relations (6.13) and $F_{10} = n_1 + n_2 + n_3 - N_f$. The invariants $\zeta_i^{(r)}$ are selected in such a way that the 10×5 matrix

$$\frac{\partial(F_i)}{\partial(\zeta_j^{(r)})} \tag{9.5}$$

has rank 5. Then $\zeta_i^{(r)}$ can be expressed in terms of $\zeta_i^{(c)}$ and N_f. There is usually more than one choice for $\zeta_i^{(c)}$, but not all choices are equally acceptable. Since we need to actually solve the equations $F_i = 0$, we must try to find a set $\zeta_i^{(c)}$ such that the solution takes a simple form. This can be done only by inspecting the solutions in each case.

For example, for the $A^{(4)}$ point $(1, 0, 0)$ with stabilizer $D_{2d}^{(x)} \times \mathcal{T}$, we find that $\zeta^{(c)} = (s_2, s_3, t_2, t_3)$ is a suitable set of invariants. The values of these invariants on the specific RE are $\zeta^{(c)*} = (0, 0, 0, 0)$. We define the displacement vector

$$d_i^{A^{(4)}} = \zeta_i = \zeta_i^{(c)} - \zeta_i^{(c)*}, \quad i = 1, \ldots, 4.$$

Notice that near the RE we can express all the invariants in terms of the displacements ζ_i.

An important difference with regard to the previous discussion is that the displacements here are not necessarily canonically conjugate variables. It is therefore important to calculate the linearized Poisson structure near each RE. In order to do this we calculate each Poisson bracket $\{\zeta_i, \zeta_j\}$ using the invariants, and then we express the result as a function of the

[48] Notation as in Table 21 with $n \equiv N_f$.

Table 33

Dynamically invariant local coordinates at the RE on the F_2-mode space $\mathbb{C}P^2$.

Type of RE		Poisson algebra[49]
$A^{(4)}$	$D_{2d}^{(x)} \times \mathcal{T}$	$\begin{array}{r\|ccc} & \zeta_2 & \zeta_3 & \zeta_4 \\ \hline s_2 = \zeta_1 & 0 & -2n & 0 \\ s_3 = \zeta_2 & & 0 & 2n \\ t_2 = \zeta_3 & & & 0 \\ t_3 = \zeta_4 & & & \end{array}$
$A^{(2)}$	$C_{2v}^{(z)} \times \mathcal{T}$	$\begin{array}{r\|ccc} & \zeta_2 & \zeta_3 & \zeta_4 \\ \hline n_2 - \frac{1}{2}n = \zeta_1 & 0 & 0 & n \\ s_1 = \zeta_2 & & n & 0 \\ t_1 = \zeta_3 & & & 0 \\ t_3 = \zeta_4 & & & \end{array}$
$A^{(3)}$	$C_{2v}^{(z)} \times \mathcal{T}$	$\begin{array}{r\|ccc} & \zeta_2 & \zeta_3 & \zeta_4 \\ \hline s_2 - 2n/3 = \zeta_1 & 0 & 0 & -2n/3 \\ s_3 - 2n/3 = \zeta_2 & & 2n/3 & 0 \\ t_2 = \zeta_3 & & & 0 \\ t_3 = \zeta_4 & & & \end{array}$
$B^{(3)}$	$C_3^{[111]} \wedge \mathcal{T}_s^{\|}$	$\begin{array}{r\|ccc} & \zeta_2 & \zeta_3 & \zeta_4 \\ \hline s_2 + n/3 = \zeta_1 & -n/\sqrt{3} & 0 & n/3 \\ s_3 + n/3 = \zeta_2 & & -n/3 & 0 \\ t_2 - n/\sqrt{3} = \zeta_3 & & & n/\sqrt{3} \\ t_3 - n/\sqrt{3} = \zeta_4 & & & \end{array}$
$B^{(4)}$	$S_4^{(x)} \wedge \mathcal{T}_2^{(y)}$	$\begin{array}{r\|ccc} & \zeta_2 & \zeta_3 & \zeta_4 \\ \hline n_3 - \frac{1}{2}n = \zeta_1 & -n & 0 & 0 \\ s_1 = \zeta_2 & & 0 & 0 \\ s_2 = \zeta_3 & & & n \\ t_2 = \zeta_4 & & & \end{array}$

displacements ζ_i, keeping terms only up to first order. We can follow the above program for all the RE. The results are summarized in Table 33. Observe that for most cases in this table it is immediately obvious how to define standard canonically conjugate variables. Thus in the case of the $A^{(3)}$ point, we can define canonically conjugate variables (ξ, η),

$$\xi_1 = \alpha\zeta_2,\ \eta_1 = \alpha\zeta_3,\ \xi_2 = \alpha\zeta_4,\ \eta_2 = \alpha\zeta_1, \quad \alpha = \sqrt{\tfrac{3}{2N_f}},$$

such that the local 2-form is $d\xi_1 \wedge d\eta_1 + d\xi_2 \wedge d\eta_2$. The only case where the proper definition of the canonical variables is not obvious is the case of the $B^{(3)}$ point.

To conclude this section, we should add one important remark. The above canonical (or noncanonical) variables can *only* be used to study *linear* Hamiltonian equations of the reduced system near the RE. This limitation is due to the fact that the symplectic form near the RE has a standard matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ only to the first order (in the limit of linearization). When nonlinear equations of motion near the RE are sought (e.g., when bifurcations of the RE are studied) this form should be further “flattened” in higher orders; see section V.8.2 of [11].

9.3. Dynamics on invariant subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. In section 8.4 and Table 28 we describe three possible types of subspaces of the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ that

[49] Notation as in Table 21 with $n \equiv N_f$.

are invariant with regard to the (symplectic) action of the spatial group T_d and are therefore dynamically invariant subspaces of the reduced system. These subspaces are either 2-spheres $\mathbb{S}^2 \sim \mathbb{C}P^1$ or a product of 2-spheres $\mathbb{S}^2 \times \mathbb{S}^2$. Dynamics on each $\mathbb{S}^2$ can be described by Euler–Poisson equations. The corresponding "angular momentum" algebras so(3) are constructed below.

The construction is most straightforward in the case of the C_2-invariant $\mathbb{S}^2$ subspace of $\mathbb{C}P^2$ (which can be represented using complex coordinates (z_1, z_2) and $z_3 = 0$). We can see from Tables 21 and 28 that the three so(3) components can be chosen as

$$(Y_1, Y_2, Y_3) = \left(\frac{s_3}{2}, \frac{t_3}{2}, \frac{x_3}{2}\right).$$

Restricting the Poisson algebra in Table 31 to the C_2^z-invariant sphere defined in Table 28 shows that this is indeed so(3),

$$\{Y_\alpha, Y_\beta\} = \epsilon_{\alpha\beta\gamma} Y_\gamma,$$

with Casimir

$$Y_1^2 + Y_2^2 + Y_3^2 = \left(\frac{N_f}{2}\right)^2.$$

Dynamics on the $\mathbb{S}^2 \times \mathbb{S}^2$ invariant subspace in Table 28 is described using an so(3) × so(3) algebra. The second so(3) is generated by (v_1, v_2, v_3), commutes with (Y_1, Y_2, Y_3), and has the Casimir

$$v_1^2 + v_2^2 + v_3^2 = \left(\frac{N_e}{2}\right)^2.$$

Dynamical variables for the C_s^{xy}-invariant sphere in Table 28 are obtained analogously. Restricting the Poisson algebra in Table 31, we find polynomials

$$\begin{aligned}(X_1, X_2, X_3) &= \left(\frac{t_1 - t_2}{2\sqrt{2}}, \frac{s_1 + s_2}{2\sqrt{2}}, \frac{2s_3 - N_f}{2}\right) \\ &= \left(\frac{\tau}{\sqrt{2}}, \frac{\sigma}{\sqrt{2}}, \frac{2\eta - 1}{2}\right) N_f,\end{aligned}$$

which form the so(3) algebra, such that

$$X_1^2 + X_2^2 + X_3^2 = \left(\frac{N_f}{2}\right)^2, \quad \{X_i, X_j\} = \epsilon_{ijk} X_k.$$

10. Existence and stability of RE.

RE are special stationary solutions of the equations of motion (9.1), which in many cases are defined entirely by symmetry and exist for any generic small symmetry-preserving perturbation. Isolated fixed points of the $T_d \times \mathcal{T}$ group action on the reduced phase space $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ found in section 8 are necessarily stationary solutions of (9.1) and are therefore representing RE. We show how to determine the stability of these RE using the reduced Hamiltonian H_{eff} defined as a function on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$. Subsequently, we search for other RE, which are not fixed by the action of the finite symmetry group.

10.1. Linear stability of fixed RE. Fixed RE of our system correspond to fixed points of the group action found in sections 8.1 and 8.2 (see Table 10, 15, and 27). Stability of these RE was already analyzed in section 5.1.4, where we studied analytical Poincaré surfaces of section using initial phase space variables $(z, \bar{z})$. Here we use invariants in Table 21 to study RE stability directly on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$, i.e., *without* lifting back to the initial phase space. As before, we can explain our approach in the example of axial symmetry with axis C_k oriented as C_4^z (see sections 8.1 and 9.2).

Analysis on $\mathbb{C}P^2$ is the most difficult [97]. Using canonical variables of the linearization limit found in section 9.2 and Table 32, we come to the problem of determining linear stability of a stationary point in different canonical planes. Each such plane has its origin at the relative equilibrium, and the stabilizer C_k of the RE acts as a rotation $C_{k'}$ about the origin. In general, the actions of C_k on the initial 3-space and on the particular canonical plane can differ. We have shown in section 7 that these actions are the same, and $k = k'$ for all planes except (x_3, s_3), where $k' = k/2$.

In the case of C_3 and C_k with $k > 4$, canonical coordinate–momentum pairs in each of the four symplectic planes transform according to a pair of conjugate complex representations of the symmetry group (which correspond to two representations of the SO(2) group of indexes $\pm m$). Variables x_3 and s_3 in the case of C_4 and all variables in the case of C_2 transform according to different real one-dimensional irreducible representations.

In order to take into account the full symmetry group $T_d \times \mathcal{T}$, as in section 5.1.5 we should find the action of the stabilizer of each RE on the local variables ζ_i, defined in section 9.2. As before, we need to define the action of the elements R of the stabilizer G of each RE on the displacement vector $d = \{\zeta_i\}_{i=1,\dots,4}$. For this we act with R on the original complex variables (z_1, z_2, z_3). This action induces a linear action L_R on the invariants, and a nonlinear action N_R on the displacements ζ_i. The last action is computed by expressing ζ_i in terms of the invariants $\zeta_i^{(r)}$, acting on them with L_R, and expressing the result again in terms of the displacements ζ_i. The action $N_R^{(1)}$ is then defined as the linearization of N_R.

The results are presented in Table 34. We can find from this table that the reducible representation of each stabilizer spanned by the variables ζ_i is decomposed into exactly the same irreducible representations as the representation of the stabilizer spanned by the local variables (x_1, x_2, y_1, y_2) of section 5.1.5.

10.2. Finding additional stationary points on invariant subspaces. In sections 8.4 and 9.3 and in Tables 28 and 30 we describe subspaces of $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ whose stabilizers are purely spatial subgroups of $T_d \times \mathcal{T}$. Such subgroups are symplectic and the subspaces are dynamically invariant subspaces of the reduced system. To find equations of motion on these subspaces we can simply restrict the reduced Hamiltonian H_{eff} using Table 28 and express it in terms of dynamical variables (Y_1, Y_2, Y_3) or (X_1, X_2, X_3), defined in Table 33, and/or (v_1, v_2, v_3). Equations of motion then can be generated using the respective so(3) algebras. Before a general solution is attempted, resulting equations first can be restricted to the one-dimensional strata of the residual group action (see section 8.4). All stationary points found should satisfy Morse conditions for the respective subspace.

As explained in section 5, we should study the C_s-invariant subspace, which is a 2-sphere with no critical orbits of the symmetry group action (no RE fixed by symmetry). This sphere

Table 34

Action of stabilizers on dynamically invariant local coordinates on $\mathbb{C}P^2$ defined in Table 33.

Action of $D_{2d}^{(x)} \times \mathcal{T}$on $E_g \oplus E_u$ for the $A^{(4)}$ RE

R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$	R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$
E	ζ_1	ζ_2	ζ_3	ζ_4	$\mathcal{T}$	ζ_1	ζ_2	$-\zeta_3$	$-\zeta_4$
C_2^x	$-\zeta_1$	$-\zeta_2$	$-\zeta_3$	$-\zeta_4$	$C_2^x\mathcal{T}$	$-\zeta_1$	$-\zeta_2$	ζ_3	ζ_4
C_2^y	ζ_1	$-\zeta_2$	ζ_3	$-\zeta_4$	$C_2^y\mathcal{T}$	ζ_1	$-\zeta_2$	$-\zeta_3$	ζ_4
C_2^z	$-\zeta_1$	ζ_2	$-\zeta_3$	ζ_4	$C_2^z\mathcal{T}$	$-\zeta_1$	ζ_2	ζ_3	$-\zeta_4$
σ^{yz}	ζ_2	ζ_1	$-\zeta_4$	$-\zeta_3$	$\sigma^{yz}\mathcal{T}$	ζ_2	ζ_1	ζ_4	ζ_3
$\sigma^{\overline{yz}}$	$-\zeta_2$	$-\zeta_1$	ζ_4	ζ_3	$\sigma^{\overline{yz}}\mathcal{T}$	$-\zeta_2$	$-\zeta_1$	$-\zeta_4$	$-\zeta_3$
S_4^x	$-\zeta_2$	ζ_1	ζ_4	$-\zeta_3$	$S_4^x\mathcal{T}$	$-\zeta_2$	ζ_1	$-\zeta_4$	ζ_3
$(S_4^x)^{-1}$	ζ_2	$-\zeta_1$	$-\zeta_4$	ζ_3	$(S_4^x)^{-1}\mathcal{T}$	ζ_2	$-\zeta_1$	ζ_4	$-\zeta_3$

Action of $S_4^{(x)} \wedge \mathcal{T}_2^{(y)}$ on $B_1 \oplus B_2 \oplus E$ for the $B^{(4)}$ RE

R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$	R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$
E	ζ_1	ζ_2	ζ_3	ζ_4	$\sigma^{yz}\mathcal{T}$	$-\zeta_1$	ζ_2	$-\zeta_4$	$-\zeta_3$
C_2^x	ζ_1	ζ_2	$-\zeta_3$	$-\zeta_4$	$\sigma^{\overline{yz}}\mathcal{T}$	$-\zeta_1$	ζ_2	ζ_4	ζ_3
S_4^x	$-\zeta_1$	$-\zeta_2$	ζ_4	$-\zeta_3$	$C_2^y\mathcal{T}$	ζ_1	$-\zeta_2$	ζ_3	$-\zeta_4$
$(S_4^x)^{-1}$	$-\zeta_1$	$-\zeta_2$	$-\zeta_4$	ζ_3	$C_2^z\mathcal{T}$	ζ_1	$-\zeta_2$	$-\zeta_3$	ζ_4

Action of $C_{2v}^{(z)} \times \mathcal{T}$ on $A_{2g} \oplus A_{2u} \oplus B_{1g} \oplus B_{1u}$ for the $A^{(2)}$ RE

R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$	R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$
E	ζ_1	ζ_2	ζ_3	ζ_4	$\mathcal{T}$	ζ_1	ζ_2	$-\zeta_3$	$-\zeta_4$
C_2^z	ζ_1	$-\zeta_2$	$-\zeta_3$	ζ_4	$C_2^z\mathcal{T}$	ζ_1	$-\zeta_2$	ζ_3	$-\zeta_4$
σ^{xy}	$-\zeta_1$	ζ_2	ζ_3	$-\zeta_4$	$\sigma^{xy}\mathcal{T}$	$-\zeta_1$	ζ_2	$-\zeta_3$	ζ_4
$\sigma^{\overline{xy}}$	$-\zeta_1$	$-\zeta_2$	$-\zeta_3$	$-\zeta_4$	$\sigma^{\overline{xy}}\mathcal{T}$	$-\zeta_1$	$-\zeta_2$	ζ_3	ζ_4

Action of $C_{3v}^{[111]} \times \mathcal{T}$ on $E_g \oplus E_u$ for the $A^{(3)}$ RE

R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$
E	ζ_1	ζ_2	ζ_3	ζ_4
$C_3^{[111]}$	$-\zeta_1 - \zeta_2$	ζ_1	$-\zeta_3 - \zeta_4$	ζ_3
$(C_3^{[111]})^2$	ζ_2	$-\zeta_1 - \zeta_2$	ζ_4	$-\zeta_3 - \zeta_4$
σ^{yz}	ζ_2	ζ_1	$-\zeta_4$	$-\zeta_3$
σ^{zx}	ζ_1	$-\zeta_1 - \zeta_2$	$-\zeta_3$	$\zeta_3 + \zeta_4$
σ^{xy}	$-\zeta_1 - \zeta_2$	ζ_2	$\zeta_3 + \zeta_4$	$-\zeta_4$
$\mathcal{T}$	ζ_1	ζ_2	$-\zeta_3$	$-\zeta_4$
$C_3^{[111]}\mathcal{T}$	$-\zeta_1 - \zeta_2$	ζ_1	$\zeta_3 + \zeta_4$	$-\zeta_3$
$(C_3^{[111]})^2\mathcal{T}$	ζ_2	$-\zeta_1 - \zeta_2$	$-\zeta_4$	$\zeta_3 + \zeta_4$
$\sigma^{yz}\mathcal{T}$	ζ_2	ζ_1	ζ_4	ζ_3
$\sigma^{zx}\mathcal{T}$	ζ_1	$-\zeta_1 - \zeta_2$	ζ_3	$-\zeta_3 - \zeta_4$
$\sigma^{xy}\mathcal{T}$	$-\zeta_1 - \zeta_2$	ζ_2	$-\zeta_3 - \zeta_4$	ζ_4

Action of $C_3^{[111]} \wedge \mathcal{T}_s^{\parallel}$ on $E \oplus E$ for the $B^{(3)}$ RE

R	$R\zeta_1$	$R\zeta_2$	$R\zeta_3$	$R\zeta_4$
E	ζ_1	ζ_2	ζ_3	ζ_4
$C_3^{[111]}$	$-\zeta_1 - \zeta_2$	ζ_1	$-\zeta_3 - \zeta_4$	ζ_3
$(C_3^{[111]})^2$	ζ_2	$-\zeta_1 - \zeta_2$	ζ_4	$-\zeta_3 - \zeta_4$
$\sigma^{yz}\mathcal{T}$	ζ_2	ζ_1	ζ_4	ζ_3
$\sigma^{zx}\mathcal{T}$	ζ_1	$-\zeta_1 - \zeta_2$	ζ_3	$-\zeta_3 - \zeta_4$
$\sigma^{xy}\mathcal{T}$	$-\zeta_1 - \zeta_2$	ζ_2	$-\zeta_3 - \zeta_4$	ζ_4

$\mathbb{S}^2$ has an invariant circle $\mathbb{S}^1$ with stabilizer $\mathcal{T}_2$, which is defined by $X_2 = 0$ (or equivalently $s_1 = s_2 = \sigma = 0$). The Morse requirements of two stationary points, a minimum and a maximum, for both the sphere $\mathbb{S}^2$ and its invariant subspace $\mathbb{S}^1$ can be satisfied if the two points lie on $\mathbb{S}^1$. To find these points we first restrict H_{eff} to the C_s sphere using Table 28 and reexpress it as a function of dynamical variables (X_1, X_2, X_3). If $H_{\text{eff}}(X)$ is a Morse function, we will always find at least two stationary points of $H_{\text{eff}}(X)$ on the $\mathcal{T}_2$-invariant circle $\mathbb{S}^1$. Such points are particular solutions to the Euler–Poisson equations

$$\dot{X} = \begin{pmatrix} 0 & X_3 & -X_2 \\ -X_3 & 0 & X_1 \\ X_2 & -X_1 & 0 \end{pmatrix}^T \nabla_X H_{\text{eff}}(X) = 0, \tag{10.1}$$

where we should set $X_2 = 0$.

11. Examples.

11.1. Vibrational structure of the F_2-mode polyads.

Possible configurations and different types of stability of the RE of the F_2-mode subsystem were studied in section 5.3.2. In section 5.2.3 we emphasized the difference between simplest and nonsimplest Hamiltonians on $\mathbb{C}P^2$ and described the system of RE in each case. Below we study this system in more detail on a model example, which is discussed in more detail in [98].

The Hamiltonian of a molecule, which has the F_2 mode, can be written as $\omega H_{\text{vib}}^{\nu_3} + H'$, where to order ϵ^2

$$\begin{aligned} H_{\text{vib}}^{\nu_3} =& \tfrac{1}{2}(q_1^2 + p_1^2) + \tfrac{1}{2}(q_2^2 + p_2^2) + \tfrac{1}{2}(q_3^2 + p_3^2) + \epsilon K_3 q_1 q_2 q_3 + \epsilon^2 K_t \tfrac{1}{2}(q_1^4 + q_2^4 + q_3^4) \\ & + \epsilon^2 K_s \tfrac{1}{2}(q_1^2 + q_2^2 + q_3^2)^2 + \epsilon^2 K_l \tfrac{1}{2}\left[\mathbf{p} \times \mathbf{q}\right]^2, \end{aligned} \tag{11.1a}$$

and H' represents other degrees of freedom and interaction of these degrees with the F_2-mode subsystem (cf. section 5.3.2). In order to study this subsystem, we should consider H' explicitly, normalize the Hamiltonian $\omega H_{\text{vib}}^{\nu_3} + H'$, and *then* restrict the obtained normal form on $\mathbb{C}P^2$ by setting to zero all dynamical variables of other subsystems. This approach was used in [13] for the case of the A_4 molecule, where we set to zero integrals j, N_e, and N_a, angular momenta (j_1, j_2, j_3), and E-mode vibrational variables (v_1, v_2, v_3); see Table 21. Furthermore, the simple atom–atom bond model of A_4 used in [13] gives

(11.1b)

Constant	ω	K_3	K_l	K_s	K_t
Value	$\sqrt{2}$	$3/2^{5/4}$	$2^{-5/2}\lambda$	$-5/2^{9/2}$	$7/2^{-9/2}$

.

Here the parameter λ is introduced so that the value of K_l obtained in [13] corresponds to $\lambda = 1$. In the present study we focus on the F_2-mode subsystem without taking interactions with other subsystems into account. To this end we use a simplified model, where H' is neglected *before* normalization. This model turns out to be sufficient for studying the transition from the simplest to the nonsimplest Hamiltonian on $\mathbb{C}P^2$.

The normal form of (11.1a) $H_{\text{eff}}^{\nu_3} = n + \epsilon^2 \mathcal{H}_2^{\nu_3} + \cdots$ can be expressed using invariants in Table 21. In the second order $\mathcal{H}_2^{\nu_3}$ we obtain

Term		Coefficient
ϵ^2	n^2	$\frac{1}{2}(K_3^2/24 + K_l + K_s + 3K_t/4)$
ϵ^2	$(s_1^2 + s_2^2 + s_3^2)$	$\frac{1}{2}(-K_3^2/4 - K_l + K_s/2)$
ϵ^2	$(2nn_3 - 3n_3^2 - x_3^2)$	$\frac{1}{2}(K_3^2/24 + K_l - K_s/2 - 3K_t/4)$

Taking the $T_d \times \mathcal{T}$ symmetry into account (Table 25), we can verify that $\mathcal{H}_2^{\nu_3}$ has indeed only three independent terms:

Term		Coefficient
ϵ^2	$V_{ffff}^{A_1}$	$(5K_s - 4K_l - K_3^2 + 3K_t)/4$
ϵ^2	V_{ffff}^{E}	$(K_s + K_l + K_3^2/4 + 3K_t/2)/\sqrt{2}$
ϵ^2	$V_{ffff}^{F_2}$	$(K_s + K_l - K_3^2/6)\sqrt{3}/2$

In fact, the Hamiltonian $\mathcal{H}_2^{\nu_3}$ has been long suggested by Hecht [95] as a model Hamiltonian for describing the internal structure of the F_2-mode polyads. Hecht expressed his model $\mathcal{H}_2^{\nu_3}$ in terms of

$$n^2 = V_{ffff}^{A_1} + \sqrt{2}V_{ffff}^{E} + \sqrt{3}V_{ffff}^{F_2}, \tag{11.2a}$$

$$t^2 = t_1^2 + t_2^2 + t_3^2 = [\mathbf{p} \times \mathbf{q}]^2 = n^2 - 3V_{ffff}^{A_1}, \tag{11.2b}$$

$$m = n_1^2 + n_2^2 + n_3^2 = V_{ffff}^{A_1} + \sqrt{2}V_{ffff}^{E}, \tag{11.2c}$$

where the "vibrational angular momentum" t^2 is often denoted as l^2 [95, 96]. Considering relations (6.13), we note an alternative to m or t,

$$s^2 = s_1^2 + s_2^2 + s_3^2 = 2n^2 - 2m - t^2. \tag{11.2d}$$

Hecht's representation

$$\mathcal{H}_2^{\nu_3} = c_0 n^2 + c_l t^2 + c_m m, \tag{11.3}$$

or, alternatively,

$$\mathcal{H}_2^{\nu_3} = c_0' n^2 + c_l' t^2 + c_s s^2 \tag{11.3'}$$

is particularly convenient for the order ϵ^2 classification of qualitatively different normal forms $H_{\text{eff}}^{\nu_3}$ on $\mathbb{C}P_n^2$. We remark that the term n^2 is just an additive ("scalar") energy constant (for the reduced system on $\mathbb{C}P_n^2$, i.e., within one n polyad). Neglecting this term, the energies of RE of the F_2-mode system with Hamiltonian (11.3) or (11.3′) are given below.

Type of RE		$\mathcal{H}_2 n^{-2} - c_0'$	$\mathcal{H}_2 n^{-2} - c_0$
$C_{2v} \times \mathcal{T}$	A	c_s	$\frac{1}{2}c_m$
$C_{3v} \times \mathcal{T}$	A	$\frac{4}{3}c_s$	$\frac{1}{3}c_m$
$C_3 \wedge \mathcal{T}_s$	B	$c_l' + \frac{1}{3}c_s$	$c_l + \frac{1}{3}c_m$
$D_{2d} \times \mathcal{T}$	A	0	c_m
$S_4 \wedge \mathcal{T}_2$	B	c_l'	$c_l + \frac{1}{2}c_m$
$C_s \wedge \mathcal{T}_2$		$(c_l')^2(c_l' - \frac{1}{4}c_s)^{-1}$	

It follows that reciprocal energies of RE of such system and, therefore, the structure of the vibrational polyads of the F_2 mode, are described (to the lowest order) by *one* parameter, the

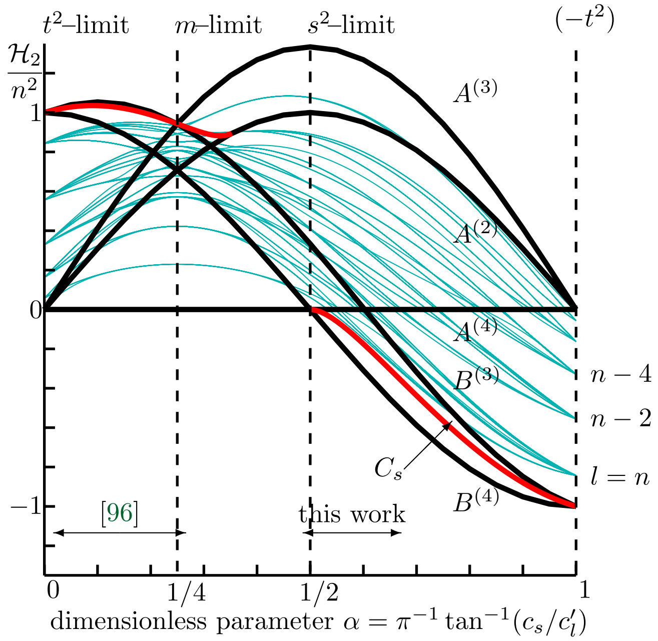


Figure 17. *Reciprocal energies (in units of n^2 and without scalar part $c'_0 n^2$) of RE (bold lines) and the structure of the $n = 10$ quantum polyad (thin blue lines) of the F_2-mode system in the second order model (Hecht's Hamiltonian (11.3′)).*

ratio of c_l and c_m or c'_l and c_s. To study all possibilities we set c'_l and c_s in (11.3′) to $\cos(\alpha\pi)$ and $\sin(\alpha\pi)$, respectively. Resulting RE energies $\mathcal{H}_2 n^{-2} - c'_0$ are shown in Figure 17.

We call the three principal limiting cases of the F_2-mode system with $\alpha = 0$, $\frac{1}{4}$, and 1 the t^2, m, and $(-t^2)$ limits, respectively (see Figure 17). All these limits have continuous symmetries and the system is integrable. The s^2 limit shown in Figure 17 is indicated primarily for convenience. It has no simple integrals and, probably, is not integrable.

The symmetry group of the t^2 limit contains the spatial group SO(3). The energy in this limit is a function of the vibrational angular momentum $t^2 = l^2$, which is zero for the A-type RE and maximum $l = n$ for the B-type RE. The corresponding quantum polyad is split into multiplets with $l = n, n-2, \ldots$. The $(-t^2)$ limit is the same as the t^2 limit, albeit for the opposite sign of energies.

The spatial symmetry of the m limit is cubic, so this limit better represents the symmetry of the system. However, this limit has continuous dynamical symmetry $T^3 = S^1 \times S^1 \times S^1$. The energy can be represented as a function of three integrals in involution (n_1, n_2, n_3), i.e., actions of individual oscillators, such that $n_1 + n_2 + n_3 = n$. The quantum states can be labeled with the three corresponding quantum numbers. The number of degenerate states is normally given by the number of permutations of the set of the three integers $[N_1, N_2, N_3]$. In the $n = 10$ example in Figure 17, it is either 3 or 6 (maximum).

Transition between the t^2 and m limits has been studied by Patterson [96], who remarked that F_2-mode polyads of certain molecules, such as SiF$_4$, SF$_6$, and UF$_6$, belong to the interval of $\alpha = [0, \frac{1}{4}]$. On the other hand, our prediction for P$_4$ suggests that the ν_3 polyads of this molecule fall into the $\alpha \geq \frac{1}{2}$ category. A complimentary approach to classifying ν_3-mode systems is by specifying the intervals of α-values, where the second order normalized

Hamiltonian (11.3′) represents particular classes of $(T_d \times \mathcal{T})$-invariant Morse functions on $\mathbb{C}P^2$ introduced in section 5.2.3. We can see in Figure 17 that there are at least three such regions. Furthermore, we notice that the $\mathcal{H}_2^{\nu_3}$ Hamiltonian is of the simplest Morse type only on the interval $\alpha = (\pi^{-1}\tan^{-1} 2, \frac{1}{2})$, where the additional nonfixed RE of symmetry $C_s \wedge \mathcal{T}_2$ does not exist.

In this paper we would like to uncover the precise role of the vibrational angular momentum term $(\mathbf{p} \times \mathbf{q})^2$ in the transition between the simplest and nonsimplest RE structures. To this end we detail our example in section 5.3.2. Note that even though $(\mathbf{p} \times \mathbf{q})^2$ enters in both the initial Hamiltonian (11.1a) and the normalized (i.e., n-polyad) Hamiltonian (11.3′), the actual vibrational angular momentum of the system is measured by the parameter K_l of (11.1a) and not by the effective parameter c_l or c_l'. Therefore, we should take the vibrational Hamiltonian (11.1a), replace K_l with λK_l, use constants (11.1b) obtained in [13], and make λ vary between, for example, 0 and 2. Normalizing and using the (s^2, t^2, n) representation (11.3′) with parameters (11.1b) gives

$$\frac{c_s}{c_l'} = \frac{5K_3^2 + 18K_t}{18K_t + 12K_s - K_3^2 - 24\lambda K_l} = \frac{51}{5 - 16\lambda}.$$

It follows that we study the F_2-mode system near the s^2 limit in the range $\alpha \approx [0.469, 0.655]$; see Figure 17.

The stability analysis of RE for $\lambda = 0$, i.e., *without* the vibrational angular momentum term $(\mathbf{p} \times \mathbf{q})^2$ in (11.1a), shows that the corresponding normal form $H_{\text{eff}}^{\nu_3}$ is a Morse function on $\mathbb{C}P^2$ of the simplest kind. In the A$_4$ molecule model of [13] (with $\lambda = 1$ and $\alpha \approx 0.567$), $H_{\text{eff}}^{\nu_3}$ is of the nonsimplest kind. The transition from the simplest to the nonsimplest case is clearly related to the $(\mathbf{p} \times \mathbf{q})^2$ term. The RE energy computed using the normal form of the Hamiltonian (11.1) as a function of the parameter λ is shown in Figure 18, where we also indicate the Morse signatures and stability types of the RE. As can be seen in this figure, the B-type RE, which are shaped as loops in the configuration space (see Figure 16) and thus induce the maximum vibrational angular momentum, respond largely to the change of λ, while the energy of the A-type RE remains unchanged. As a consequence, the RE structure as a whole has to change qualitatively. This change involves two bifurcations.

The first bifurcation happens at $\lambda = 5/16$ $(\alpha = 0)$. The $B^{(4)}$ relative equilibrium, which was unstable with Morse index 1 (Poincaré index +2) for $\lambda < 5/16$, becomes stable with index 0 (+4) for $\lambda > 5/16$. At the moment of bifurcation the energies of the $B^{(4)}$ and $A^{(4)}$ RE are equal, then as λ increases, the energy of $B^{(4)}$ becomes greater. At the same time, the $A^{(4)}$ relative equilibrium, which was stable with Morse index 0 (+4), becomes doubly unstable with Morse index 2 (0). In order for the Morse conditions to be satisfied globally, a *new* relative equilibrium bifurcates from $A^{(4)}$. This is the $C_s \wedge \mathcal{T}_2$ symmetric RE described in section 5.2.3. The new relative equilibrium is unstable with Morse index 2 (0). Part of the described bifurcation can be regarded as a so-called "pitchfork bifurcation," or a bifurcation with broken symmetry Z_2. Indeed, the system restricted to the C_s sphere (see section 5.2.3) undergoes such bifurcation. Taken to the whole four-dimensional space, this phenomenon is more complex because it involves the $B^{(4)}$ relative equilibrium.

The second bifurcation happens at $\lambda = 11/8$ $(\alpha \approx 0.1)$. The moment of bifurcation is easy to notice because the second normal form energies of the $B^{(3)}$ and $A^{(4)}$ RE (i.e., values of the

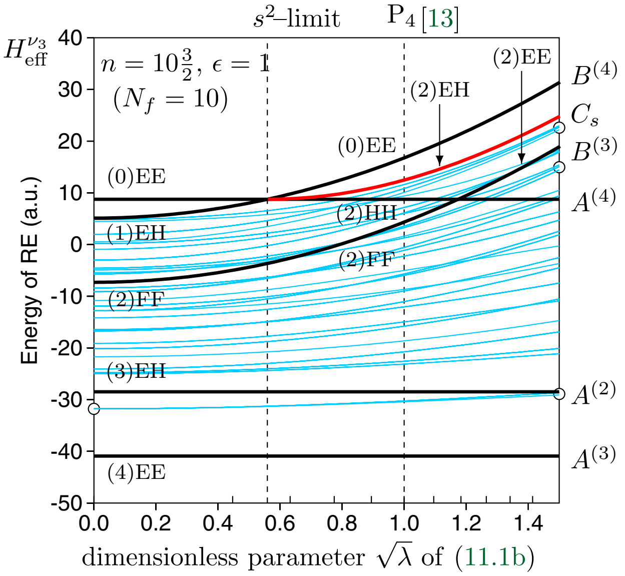


Figure 18. *Quantum (grey lines) and classical RE (bold lines) correlation diagram for the F_2-mode system described by the one-parameter family of Hamiltonians (11.1) with fixed $\epsilon = 1$ and classical action $n = 10\frac{3}{2}$. Quantum polyads are computed for $N_f = 10$. Morse index and Hamiltonian stability are indicated for each RE; EE, EH, HH, and FF stand for elliptic–elliptic (stable), elliptic–hyperbolic (unstable), hyperbolic–hyperbolic (doubly unstable), and focus–focus (complex unstable), respectively. Circles mark level clusters discussed in the text. To compare to Figure 17, inverse the energy axis.*

Hamiltonian (11.3′) shown in Figures 18 and 17) become equal. The $B^{(3)}$ relative equilibrium undergoes a Hamiltonian Hopf bifurcation [97, 98], and from a focus–focus (complex unstable) relative equilibrium at $\lambda < 11/8$ it becomes elliptic (linearly stable) at $\lambda > 11/8$. The Morse index does not change. Unlike the previous case, the Morse requirements remain satisfied globally and there is no need for changing the number and/or stability of other RE.

Information on the RE of the system with Hamiltonian (11.1) can be used to characterize the spectrum of the corresponding quantum system as proposed in section 5.4.3. We consider the RE energies as functions of the action n and compare them to the quantum energy levels (see Figure 19). Levels with the same quantum number N_f form a *polyad* whose structure can be related to the reciprocal RE energies for corresponding classical action $n = N_f + \frac{3}{2}$. Like RE, polyads are described using the normalized Hamiltonian $H_{\text{eff}}^{\nu_3}$, where we can distinguish between "scalar" and "splitting" terms. The former depend only on n and describe an average increase in energy; the latter describe the internal structure of polyads. In the simplest approximation given by the second order normal form (11.3′), or the Hecht Hamiltonian, the internal structure of polyads is described by one-parameter α; see Figure 18.

Provided that the model potential of the P$_4$ molecule [13] is qualitatively correct, the ν_3 polyads of P$_4$ should correspond to the value of $\alpha \approx 0.6$ near the so-called s^2 limit. The most characteristic feature of the ν_3 polyads with such α is the presence of level clusters at the $A^{(3)}$ end (maximum in Figure 17 and minimum in Figures 18 and 19). The limiting $A^{(3)}$ cluster has four levels and in the case of $N_f = 10$ decomposes into symmetry components $A + F$ (Figure 18). At higher $N_f = 15$ (Figure 19) we can even see the second cluster of eight

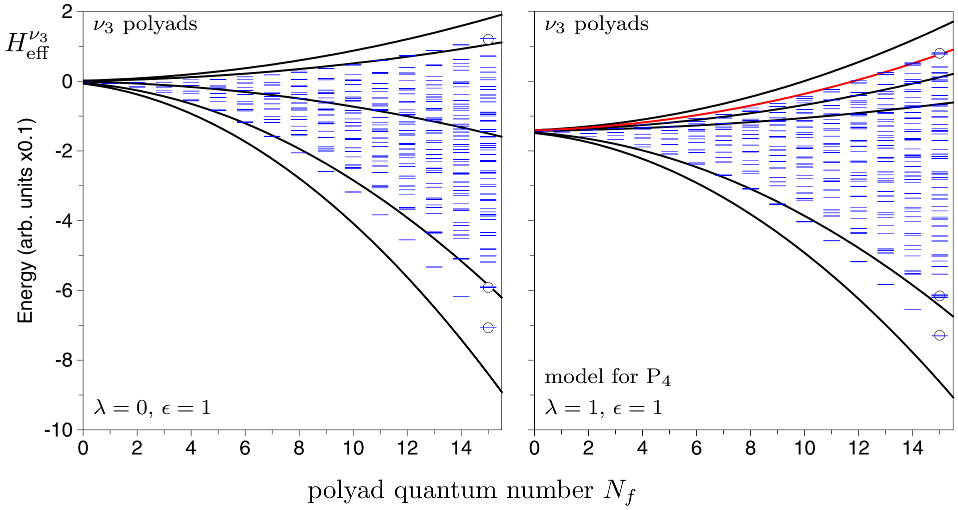


Figure 19. *Spectrum of quantum levels and energies of the RE of the F_2-mode system with Hamiltonian (11.1) with $\epsilon = 1$ and $\lambda = 0$ (left), $\lambda = 1$ (right). Circles mark $N_f = 15$ level clusters discussed in the text, and classical RE energy is plotted for $n = N_f + \frac{3}{2}$.*

levels with components $F + F + E$. As can be seen in Figures 18 and 19, the $A^{(3)}$ clusters remain insensitive to large variations of the structure parameter α or λ. The situation is more unclear at the opposite energy end of the ν_3 polyads. If, as predicted, $\alpha > \frac{1}{2}$, then the $B^{(4)}$ clusters of six levels should appear as shown in Figure 18 (top right) and 19 (topmost level of the $N_f = 10, \ldots, 15$ polyads). They decompose as $A + F + E$ for $N_f = 10$ and $F + F$ for $N_f = 15$. If, however, α for P$_4$ turns out to be sufficiently smaller than $\frac{1}{2}$, then we should expect $A^{(4)}$ clusters of three levels (such as the lowest energy levels of the $N_f = 10$ polyad for $\alpha = 0.4$ in Figure 17). Furthermore, if the $B^{(3)}$ relative equilibrium becomes sufficiently stable, a corresponding eight-fold cluster might also show up.

Several aspects should be taken into account in order to continue our analysis of the ν_3 polyad structure presented in this paper. A simple analysis based on energy separation fails as different systems of localized states overlap, and complimentary information on expectation values of characteristic dynamical invariants should be used. Degeneracy of quantum states caused by symmetry can either enhance or obscure the presence of level clusters. The position of the limiting localized state and the corresponding RE depends on the stability of the RE and the relation between n and N_f.

11.2. Rotational structure of the F_2-mode polyads. Consider an effective Hamiltonian H_{eff} commonly used to describe rotational structure of low excited F_2-mode vibrations. In the spectroscopic notation of Table 25, this Hamiltonian can be written as

$$H_{\text{eff}} = \omega_f n_f + Bj^2 - Dj^4 - h_{ff}^{2(0,A_1)} \frac{4}{3} n_f j^2 \tag{11.4a}$$

$$+ h_{ff}^{1(1,F_1)} H_{ff}^{1(1,F_1)} + \sum_{\Gamma=E,F_2} h_{ff}^{2(2,\Gamma)} H_{ff}^{2(2,\Gamma)} \tag{11.4b}$$

$$+\sum_{K=1,3} h_{ff}^{3(K,F_1)} H_{ff}^{3(K,F_1)} - \frac{\sqrt{30}}{8} D_t H^{4(4,A_1)}, \tag{11.4c}$$

where ω_f is harmonic frequency of the F_2 mode, B is the rotational constant of the molecule (for a tetrahedral molecule A_4 with four atoms of mass m whose equilibrium positions lie at a distance R from the center of mass, the constant B equals $1/(2mR^2)$), and h_{ff}, D, and D_t are parameters of higher order terms in the reduced Hamiltonian. We can omit the terms (11.4a), which have constant value in the reduced system. The energies of fixed RE can be found straightforwardly as values of the H_{eff} in (11.4) at the points listed in Table 27.

To find the two remaining RE, we restrict H_{eff} in (11.4) to the C_s-invariant sphere using the definition of this sphere in Table 28 and express the result as a function of dynamical variables X of the C_s restricted system,

$$H_{\text{eff}}^{C_s} = b(J)X_1 - a(J)X_3 + c(J, N_f), \tag{11.5}$$

where

$$\begin{aligned} a &= \sqrt{2}\, h_{ff}^{2(2,E)} J^2 - \frac{2}{\sqrt{3}}\, h_{ff}^{2(2,F_2)} J^2, \\ b &= -\frac{2\sqrt{2}}{\sqrt{3}}\, h_{ff}^{1(1,F_1)} J + \left[\frac{8\sqrt{2}}{\sqrt{3}}\, h_{ff}^{3(1,F_1)} + \frac{4}{\sqrt{5}}\, h_{ff}^{3(3,F_1)}\right] \frac{J^3}{\sqrt{3}}, \\ c &= D_t J^4 + \frac{\sqrt{2}}{6}\, h_{ff}^{2(2,E)} N_f J^2 + \frac{1}{\sqrt{3}}\, h_{ff}^{2(2,F_2)} N_f J^2. \end{aligned}$$

Note that $H_{\text{eff}}^{C_s}$ is linear in X and that it is invariant with regard to $T_d \times \mathcal{T}$ and to its subgroup $\mathcal{T}_2$ and therefore cannot depend linearly on X_2. Equations of motion (10.1) for this Hamiltonian are very simple:

$$\dot{X} = (-aX_2, bX_3 + aX_1, -bX_2).$$

Setting $X_2 = 0$ in these equations gives the condition for an equilibrium point of $H_{\text{eff}}^{C_s}$ on the $\mathcal{T}_2$-invariant circle,

$$bX_3 + aX_1 = 0,$$

which should be satisfied together with the defining equation of the circle

$$X_1^2 + X_3^2 = \frac{1}{4} N_f^2.$$

Since $a(J)$ and $b(J)$ depend differently on J; the two solutions

$$(X_1, X_2, X_3) = \pm \frac{N_f}{2\sqrt{b^2 + a^2}} (b, 0, -a)$$

move along the circle when J changes: when J is small and $b \gg a$ they are close to the point where $X_3 = 0$; at large J they approach $X_1 = 0$. The energies of these RE are

$$\pm \frac{N_f}{2} \sqrt{b(J)^2 + a(J)^2} + c(J, N_f). \tag{11.6}$$

Table 35

Energy of rotation–vibration RE in the case of low excited F_2-mode vibrations of a tetrahedral molecule A_4.

Point[50]	Energy of RE (values of H_{eff})
$A^{(2)}$	$-\frac{\sqrt{2}}{3}h_{ff}^{2(2,E)}N_fJ^2-\frac{2}{\sqrt{3}}h_{ff}^{2(2,F_2)}N_fJ^2+D_tJ^4$
$A^{(3)}$	$-\frac{8\sqrt{3}}{9}h_{ff}^{2(2,F_2)}N_fJ^2+\frac{8}{3}D_tJ^4$
$A^{(4)}$	$-\frac{4\sqrt{2}}{3}h_{ff}^{2(2,E)}N_fJ^2-4D_tJ^4$
$B,C^{(3)}$	$\mp\frac{\sqrt{2}}{\sqrt{3}}h_{ff}^{1(1,F_1)}N_fJ+\frac{4\sqrt{3}}{9}h_{ff}^{2(2,F_2)}N_fJ^2$ $\pm\frac{4\sqrt{2}}{3}\left(h_{ff}^{3(1,F_1)}+\frac{2\sqrt{2}}{\sqrt{15}}h_{ff}^{3(3,F_1)}\right)N_fJ^3+\frac{8}{3}D_tJ^4$
$B,C^{(4)}$	$\mp\frac{\sqrt{2}}{\sqrt{3}}h_{ff}^{1(1,F_1)}N_fJ+\frac{2\sqrt{2}}{3}h_{ff}^{2(2,E)}N_fJ^2$ $\pm\frac{4\sqrt{2}}{3}\left(h_{ff}^{3(1,F_1)}-\frac{\sqrt{2}}{\sqrt{15}}h_{ff}^{3(3,F_1)}\right)N_fJ^3-4D_tJ^4$

The simplest way to compare the energies of RE in Table 35 and (11.6) to molecular energy levels is to plot all of them in a form of an energy-momentum diagram for fixed vibrational integral N_f. In the lowest excited vibrational quantum state of the ν_3 mode, also called the fundamental or harmonic state, the quantum number $\hat{N}_f$, equals 1 which corresponds to the classical value $N_f = 1 + \frac{3}{2}$.

We illustrate our results using the Hamiltonian H_{eff} (without the scalar part (11.4a)), which describes the ν_3 vibration of the CH_4 molecule. Parameters of this Hamiltonian can be taken from [92].

$h_{ff}^{1(1,F_1)}$	-0.706007	D_t	4.42516×10^{-6}
$h_{ff}^{2(2,E)}$	1.5760×10^{-2}	$h_{ff}^{2(2,F_2)}$	-0.7220×10^{-2}
$h_{ff}^{3(1,F_1)}$	-0.635×10^{-4}	$h_{ff}^{3(3,F_1)}$	-0.187×10^{-4}

The quantum energy level spectrum of the $\nu_3 = 1$ state is shown schematically by the shaded area. This spectrum exhibits three characteristic branches formed due to the first order Coriolis interaction [81]. As shown in section 7.2.5 and Table 25, the term describing this interaction in the reduced system is the scalar product (t, j), which has spherical symmetry, i.e., to the first order; the energy depends on the angle between the 3-vectors t and j. Assuming $H_1 \propto (t, j)$, as in the case of CH_4, the energy is maximal, minimal, or zero when t and j are parallel (RE of type B), antiparallel (RE of type C), or orthogonal (RE of type A), respectively. Since the quantity (t, j) is (approximately) conserved, we can introduce another angular momentum $r = j - t$ and represent the first order energy as function of r^2, t^2, and j^2 called "rotational," "vibrational," and total angular momenta, respectively. In the quantum

[50] Fixed points of the $T_d \times \mathcal{T}$ group action on $\mathbb{C}P^2 \times \mathbb{C}P^1 \times \mathbb{S}^2$ defined in Table 27; signs + and − in the notation ± (equivalently, − and + in ∓) correspond to points B and C, respectively.

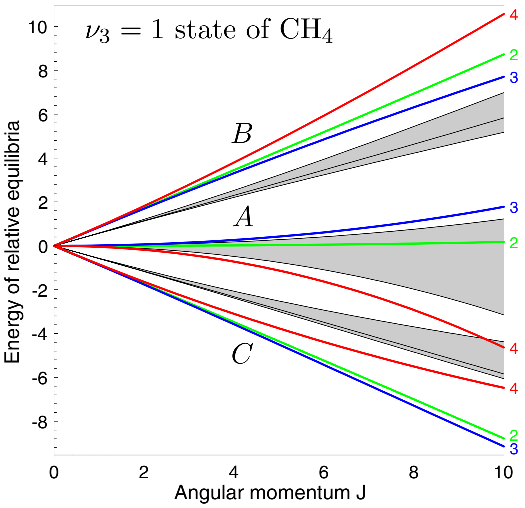


Figure 20. *Energies of rotation–vibration RE for the ν_3 mode of the CH_4 molecule. Colored lines show classical energies (see Table 35 and equation (11.6)) with N_f set to $\frac{5}{2}$; indices $(2,3,4)$ on the right give the symmetry of the corresponding RE. Thin black lines show the same energies with $N_f = 1$; shaded area represents quantum multiplets.*

state with $\hat{N}_f = 1$, the quantum number $\hat{t}$ equals 1 and the quantum number $\hat{r}$ takes the values of $\hat{\jmath}-1$, $\hat{\jmath}$, and $\hat{\jmath}+1$. These values label the three Coriolis branches of the $\nu_3 = 1$ state.

Bold lines in Figure 20 represent energies of the nine RE (two noncritical orbits are given in (11.6) and seven critical orbits in Table 35) with N_f set to its classical value $1+\frac{3}{2}$. The same energies—but with $N_f = 1$—are shown by thin lines, which border *exactly* the three rotational branches of quantum levels. This suggests a straightforward semiclassical interpretation. The RE energy with $N_f = \frac{5}{2}$ gives classical limit (classical extremum) energy for rotation–vibration levels; with $N_f = 1$ we approximate vibrational quantum energy of the state localized near the corresponding RE and obtain classical limit energy for the rotational structure only. In the case of the Hamiltonian (11.4) whose vibrational part is quadratic, this approximation matches exactly the extrema of the so-called "rotational energy surfaces" [32, 36, 35], which are obtained when all rotational operators in H_{eff} are replaced by their quantum analogues. Further examples can be found in [58, 13, 12].

12. Discussion of the results. This paper, together with [13], reports on the first substantial attempt to extend the analysis of molecular energy levels based on RE (also known as nonlinear normal modes and, in some cases, local modes, principal periodic orbits, stationary axes of rotation, etc.) from simple, often model systems to complex rotation–vibration Hamiltonians of real molecules. Our predecessors (see section 1) studied classical vibrational systems with two or three degrees of freedom [14, 15, 16, 77, 70], notably a great number of triatomic molecules [19, 22, 23, 24, 25, 26, 27, 99, 100, 101, 102, 103, 104, 105, 106], and rotational systems [31, 32, 33, 34, 37, 38, 55, 28, 29, 12]. Generalization of these studies to combined systems led to "hybrid" quantum classical systems [36, 35, 41, 42, 43, 56, 107, 108, 109]. We take the next step by studying the whole of the combined system classically and using the results for the interpretation and prediction of the corresponding quantum system. We consider the example of the rotating tetahedral molecule A_4 with six internal vibrational degrees of freedom, a system which is, arguably, at the limit of molecular systems whose rotation–vibration energy levels have already been studied in detail.

Our principal molecular result is the relation of rotation–vibration RE and the structure of the rotation–vibration energy level spectrum. Thus, we show how extremal quantum states in the rotation–vibration multiplet are associated with particular periodic rotation–vibration motion of the molecule. We took advantage of the simplicity of the classical RE description in order to analyze the structure of highly excited energy levels in different limits. In particular, we compared the structure of rotationally excited polyads to that in the case of high purely vibrational excitation. We found that when the interaction of the rotational and vibrational subsystems is significant, RE become qualitatively different from what can be expected for (or deduced from) the separable system. We predicted qualitative modifications of the system of RE and then followed it with a concrete example. This is our main mathematical result.

We also took advantage of the rich topological structure and high symmetry of our example system in order to predict and explain many important basic qualitative features of this complex system. Subsequently, we confirmed our predictions quantitatively. In particular we analyzed existence and stability of rotation–vibration RE. We extend this study to different parametric limits of molecular potential and corresponding limiting cases of the normalized system (polyads). Our more specialized mathematical results concern group-theoretical aspects of combining two subsystems, in particular the analysis of the group action on the combined phase space on the basis of the action on individual subspaces. The last, but not least, is the dynamically invariant formulation of the theory in section 9, which is the weapon of choice for further analysis of the hidden regular structures of seemingly irregular highly excited molecular states.

REFERENCES

[1] E. B. Wilson, J. C. Decius, and P. C. Cross, *Molecular Vibrations*, McGraw–Hill, New York, 1955.

[2] G. Amat, H. H. Nielsen, and G. Tarrago, *Rotation-Vibrations of Polyatomic Molecules*, Marcel Dekker, New York, 1971.

[3] J. K. G. Watson, *Higher-degree centrifugal distortion of tetrahedral molecules*, J. Mol. Spectrosc., 55 (1975), pp. 498–499.

[4] M. R. Aliev and J. K. G. Watson, *Higher-order effects in the vibration-rotation spectra of semirigid molecules*, in Molecular Spectroscopy: Modern Research, Vol. 3, K. N. Rao, ed., Academic Press, New York, 1985, pp. 1–67.

[5] Vl. I. Arnol'd, *Mathematical Methods of Classical Mechanics*, 2nd ed., Grad. Texts in Math. 60, Springer-Verlag, Berlin, 1989.

[6] Vl. I. Arnol'd, *Geometrical Methods in the Theory of Ordinary Differential Equations*, 2nd ed., Fund. Principles Math. Sci. 250, Springer-Verlag, New York, 1988.

[7] Vl. I. Arnol'd, V. V. Kozlov, and A. I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics: Dynamical Systems* III, Encyclopaedia Math. Sci. 3, Springer-Verlag, Berlin, 1993. Reprinted 1997

[8] J. M. Souriau, *Structure des Systèmes Dynamiques*, Dunod, Paris, 1970.

[9] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.

[10] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison–Wesley, Reading, MA, 1978.

[11] R. H. Cushman and L. M. Bates, *Global Aspects of Classical Integrable Systems*, Birkhäuser, Basel, 1997.

[12] Ch. van Hecke, D. A. Sadovskií, and B. I. Zhilinskií, *Qualitative analysis of molecular rotation starting from inter-nuclear potential*, European Phys. J. D At. Mol. Opt. Phys., 7 (1999), pp. 199–209.

[13] Ch. van Hecke, D. A. Sadovskií, B. I. Zhilinskií, and V. Boudon, *Rotational-vibrational relative equilibria and the structure of quantum energy spectrum of the tetrahedral molecule P_4*, Eur. Phys. J. D At. Mol. Opt. Phys., 17 (2001), pp. 13–35.

[14] J. Montaldi, R. M. Roberts, and I. Stewart, *Periodic solutions near equilibria of symmetric Hamiltonian systems*, Philos. Trans. Roy. Soc. London Ser. A, 325 (1988), pp. 237–293.

[15] J. Montaldi, M. Roberts, and I. Stewart, *Existence of nonlinear normal modes of symmetric Hamiltonian systems*, Nonlinearity, 3 (1990), pp. 695–730.

[16] J. Montaldi, M. Roberts, and I. Stewart, *Stability of nonlinear normal modes of symmetric Hamiltonian systems*, Nonlinearity, 3 (1990), pp. 730–772.

[17] B. I. Zhilinskií, *Qualitative analysis of vibrational polyads: N mode case*, Chem. Phys., 137 (1989), pp. 1–13.

[18] D. A. Sadovskií and B. I. Zhilinskií, *Group theoretical and topological analysis of localized vibration-rotation states*, Phys. Rev. A, 47 (1993), pp. 2653–2671.

[19] N. Fulton, J. Tennyson, D. A. Sadovskií, and B. I. Zhilinskií, *Nonlinear normal modes and local bending vibrations of H_3^+ and D_3^+*, J. Chem. Phys., 99 (1993), pp. 906–918.

[20] R. H. Cushman and D. Rod, *Reduction of the semisimple* 1:1 *resonance*, Phys. D, 6 (1982), pp. 105–112.

[21] R. H. Cushman, *Geometry of the bifurcations of the Henon–Heiles family*, Proc. Roy. Soc. London Ser. A, 382 (1982), pp. 361–371.

[22] L. Xiao and M. E. Kellman, *Unified semiclassical dynamics for molecular resonance spectra*, J. Chem. Phys., 90 (1989), pp. 6086–6098.

[23] Z.-M. Lu and M. E. Kellman, *Phase space structure of triatomic molecules*, J. Chem. Phys., 107 (1997), pp. 1–15.

[24] L. E. Fried and G. S. Ezra, *Semi-classical quantization using classical perturbation theory: Algebraic quantization of multidimensional spectra*, J. Chem. Phys., 86 (1987), pp. 6270–6282.

[25] Ch. Jaffé, *Comment on "Semiclassical phase space evolution of Fermi resonance spectra,"* J. Chem. Phys., 89 (1988), pp. 3395–3396.

[26] M. E. Kellman, *Approximate constants of motion for vibrational spectra of many-oscillator systems with multiple anharmonic resonances*, J. Chem. Phys., 93 (1990), pp. 6630–6635.

[27] M. E. Kellman and G. Chen, *Approximate constants of motion and energy transfer pathways in highly excited acetylene*, J. Chem. Phys., 95 (1991), pp. 8671–8672.

[28] J. Montaldi, *Persistence and stability of relative equilibria*, Nonlinearity, 10 (1997), pp. 449–466.

[29] J. Montaldi and R. M. Roberts, *Relative equilibria of molecules*, J. Nonlinear Sci., 9 (1999), pp. 53–88.

[30] I. N. Kozin, D. A. Sadovskií, and B. I. Zhilinskií, *Assigning Vibrational Polyads Using Relative Equilibria. Application to Ozone*, in preparation.

[31] A. Dorney and J. K. G. Watson, *Forbidden rotational spectra of polyatomic molecules. Stark effect and $\Delta J = 0$ transitions of T_d molecules*, J. Mol. Spectrosc., 42 (1972), pp. 135–148.

[32] W. G. Harter and C. W. Patterson, *Orbital level splitting in octahedral symmetry and SF_6 rotational spectra*, J. Chem. Phys., 66 (1977), pp. 4872–4885.

[33] C. W. Patterson and W. G. Harter, *Orbital level splitting in octahedral symmetry and SF_6 rotational spectra* II. *Quantitative features of high J levels*, J. Chem., Phys., 66 (1977), pp. 4886–4892.

[34] W. G. Harter and C. W. Patterson, *Rotational energy surfaces and high-J eigenvalue structure of polyatomic molecules*, J. Chem. Phys., 80 (1984), pp. 4241–4261.

[35] W. G. Harter, *Computer graphical and semiclassical approaches to molecular rotations and vibrations*, Comp. Phys. Rep., 8 (1988), pp. 319–394.

[36] W. G. Harter, *Molecular symmetry and dynamics*, in Atomic, Molecular, and Optical Physics Handbook, G. W. F. Drake, ed., AIP Press, New York, 1996, pp. 378–393.

[37] B. I. Zhilinskií and I. M. Pavlichenkov, *Critical phenomena in rotational spectra*, Soviet Phys. JETP, 65 (1987), pp. 221–229.

[38] I. M. Pavlichenkov and B. I. Zhilinskií, *Critical phenomena in rotational spectra*, Ann. Phys., 184 (1988), pp. 1–32.

[39] B. I. Zhilinskií, *Topological and symmetry features of intramolecular dynamics through high-resolution molecular spectroscopy*, Spectrochim. Acta A, 52 (1996), pp. 881–900.

[40] B. I. Zhilinskií, *Teoriy͡a slož͡hnyk͡h molekuly͡arnyk͡h spektrov*, Moscow University Press, Moscow, 1989.

[41] D. A. Sadovskií and B. I. Zhilinskií, *Qualitative analysis of vibration-rotation Hamiltonians for spherical top molecules*, Molec. Phys., 65 (1988), pp. 109–128.

[42] Vl. M. Krivtsun, D. A. Sadovskií, and B. I. Zhilinskií, *Critical phenomena and diabolic points in rovibrational energy spectra of spherical top molecules*, J. Molecular Spectr., 139 (1990), pp. 126–46.

[43] D. A. Sadovskií, B. I. Zhilinskií, J.-P. Champion, and G. Pierre, *Manifestations of bifurcations and diabolic points in molecular energy spectra*, J. Chem. Phys., 92 (1990), pp. 1523–1537.

[44] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, World Scientific, Singapore, 2000.

[45] M. Golubitsky and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Vol.* 1, Springer-Verlag, Berlin, 1985.

[46] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Vol.* 2, Springer-Verlag, Berlin, 1988.

[47] V. Boudon, E. B. Mkadmi, H. Bürger, and G. Pierre, *High-resolution FTIR spectroscopy and analysis of the ν_3 fundamental band of P_4*, Chem. Phys. Lett., 305 (1999), pp. 21–27.

[48] L. Michel and B. I. Zhilinskií, *Symmetry, invariants, topology. Basic tools*, Phys. Rep., 341 (2001), pp. 11–86.

[49] L. Michel and B. I. Zhilinskií, *Rydberg states of atoms and molecules. Basic group-theoretical and topological analysis*, Phys. Rep., 341 (2001), pp. 173–264.

[50] B. I. Zhilinskií, *Symmetry, invariants, and topology in molecular models*, Phys. Rep., 341 (2001), pp. 85–172.

[51] R. Littlejohn and M. Reinsch, *Gauge fields in the separation of rotations and internal motions in the n-body problem*, Rev. Modern Phys., 69 (1997), pp. 213–275.

[52] A. Tachibana and T. Iwai, *Complete molecular Hamiltonian based on the Born–Oppenheimer adiabatic approximation*, Phys. Rev. A, 33 (1986), pp. 2262–2269.

[53] A. Guichardet, *On rotation and vibration motions of molecules*, Ann. Inst. H. Poincaré Phys. Théor, 40 (1984), pp. 329–342.

[54] D. A. Sadovskií and B. I. Zhilinskií, *Counting levels within vibrational polyads. Generating function approach*, J. Chem. Phys., 103 (1995), pp. 10520–10536.

[55] I. N. Kozin and I. M. Pavlic͡henkov, J. Chem. Phys., 104 (1996), pp. 4105–4113.

[56] G. Pierre, D. A. Sadovskií, and B. I. Zhilinskií, *Organization of quantum bifurcations: Crossover of rovibrational bands in spherical top molecules*, Europhys. Lett., 10 (1989), pp. 409–414.

[57] O. I. Davarashvili, B. I. Zhilinskií, V. M. Krivt͡sun, D. A. Sadovskií, and E. P. Snegirev, *Experimental study of the sequence of bifurcations which causes the crossover of the rotational multiplet*, Soviet JETP Lett., 51 (1990), pp. 17–19.

[58] G. Dhont, D. A. Sadovskií, B. I. Zhilinskií, and V. Boudon, *Analysis of the "unusual" vibrational components of triply degenerate vibrational mode $\boldsymbol{\nu_6}$ of Mo(CO)$_6$ based on the classical interpretation of the effective rotation-vibration Hamiltonian*, J. Molec. Spectr., 201 (2000), pp. 95–108.

[59] V. E. Pavlov-Verevkin, D. A. Sadovskií, and B. I. Zhilinskií, *On the dynamical meaning of diabolic points*, Europhys. Lett., 6 (1988), pp. 573–578.

[60] D. A. Sadovskií and B. I. Zhilinskií, *Monodromy, diabolic points, and angular momentum coupling*, Phys. Lett. A, 256 (1999), pp. 235–244.

[61] R. H. Cushman and J. J. Duistermaat, *The quantum mechanical spherical pendulum*, Bull. Amer. Math. Soc., 19 (1988), pp. 475–479.

[62] San Vū Ngọc, *Quantum monodromy in integrable systems*, Comm. Math. Phys., 203 (1999), pp. 465–479.

[63] D. A. Sadovskií and R. H. Cushman, *Monodromy in perturbed Kepler systems: Hydrogen atom in crossed fields*, Europhys. Lett., 47 (1999), pp. 1–7.

[64] D. A. Sadovskií and R. H. Cushman, *Monodromy in the hydrogen atom in crossed fields*, Phys. D, 142 (2000), pp. 166–196.

[65] M. S. Child, T. Weston, and J. Tennyson, *Quantum monodromy in the spectrum of H_2O and other systems*, Molecular Phys., 96 (1999), p. 371.

[66] L. D. Landau and E. M. Lifshitz, *Quantum Mechanics*, Pergamon Press, Oxford, UK, 1965.

[67] M. Hamermesh, *Group Theory and Its Application to Physical Problems*, Addison–Wesley, Reading, MA, 1964.

[68] L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics. Theory and Applications*, Addison–Wesley, Reading, MA, 1981.

[69] J. Schwinger, *On angular momentum*, in Quantum Theory of Angular Momentum, L. C. Biedenharn and H. van Dam, eds., Academic Press, New York, 1975, pp. 229–279.

[70] D. A. Sadovskií and B. I. Zhilinskií, *Qualitative study of a model three level Hamiltonian with $SU(3)$ dynamical symmetry*, Phys. Rev. A, 48 (1993), pp. 1035–1044.

[71] W. Gröbner, *Die Lie-Reihen und ihre Anwendungen*, Mathematische Monographien 3, VEB Deutscher Verlag der Wissenschaften, Berlin, 1960.

[72] W. Gröbner, *Contributions to the Method of Lie Series*, Bibliographisches Institut, Mannheim, 1967.

[73] A. Deprit, *Canonical transformations depending on a small parameter*, Celestial Mech., 1 (1969), pp. 12–30.

[74] J. Henrard, *On a perturbation theory using Lie transforms*, Celestial Mech., 3 (1970), p. 107–120.

[75] L. Michel, *Points critiques des fonctions invariantes sur une G-varieté*, C. R. Acad. Sci. Paris Sér. A-B, 272 (1971), pp. 433–436.

[76] L. Michel, *Symmetry defects and broken symmetry*, Rev. Modern Phys., 52 (1980), pp. 617–650.

[77] R. C. Churchill, M. Kummer, and D. L. Rod, *On averaging, reduction and symmetry in Hamiltonian systems*, J. Differential Equations, 49 (1983), pp. 359–414.

[78] D. A. Sadovskií, *Normal forms, geometry, and dynamics of atomic and molecular systems with symmetry*, in Symmetry and Perturbation Theory, D. Bambusi, M. Cadoni, and G. Gaeta, eds., World Scientific, Singapore, 2001, pp. 191–205.

[79] W. G. Harter and C. W. Patterson, *Asymptotic eigensolutions of fourth and sixth rank octahedral tensor operators*, J. Math. Phys., 20 (1979), pp. 1453–1459.

[80] B. I. Zhilinskií and S. Brodersen, *Symmetry of the vibrational components in the T_d molecules*, J. Mol. Spectrosc., 163 (1994), pp. 326–338.

[81] G. Herzberg, *Spectra and Structure of Polyatomic Molecules*, R. E. Kreiger, Malabar, FL, 1989.

[82] F. Faure and B. I. Zhilinskií, *Topological Chern indices in molecular spectra*, Phys. Rev. Lett., 85 (2000), pp. 960–963.

[83] H. Weyl, *Classical Groups. Their Invariants and Representations*, Princeton University Press, Princeton, NJ, 1939.

[84] V. E. Pavlov-Verevkin and B. I. Zhilinskií, *Effective Hamiltonians for vibrational polyads: Integrity basis approach*, Chem. Phys., 126 (1988), pp. 243–253.

[85] J. Moret-Bailly, *Introduction au calcul de l'energie de vibration-rotation des molécules à symétrie sphèrique*, Cahiers de Phys., 13 (1959), pp. 476–494.

[86] J. Moret-Bailly, *Sur l'interpretation des spectres des molécules à symètrie tetraedrique*, Cahiers de Phys., 15 (1961), pp. 237–314.

[87] J. Moret-Bailly, L. Gautier, and J. Montagutelli, *Clebsch-Gordan coefficients adapted to cubic symmetry*, J. Mol. Spectrosc., 15 (1965), pp. 355–377.

[88] F. Michelot, J. Moret-Bailly, and K. Fox, *Rotational energy for spherical tops* I. *Vibronic ground state*, J. Chem. Phys., 60 (1974), pp. 2606–2609.

[89] J.-P. Champion, G. Pierre, F. Michelot, and J. Moret-Bailly, *Composantes cubiques normales des tenseurs sphèriques*, Canad. J. Phys., 55 (1977), pp. 512–520.

[90] J.-P. Champion, *Development complet de l'hamiltonien de vibration-rotation adapté à l'étude des interactions dans les molécules toupies sphèriques*, Canad. J. Phys., 55 (1977), pp. 1802–1828.

[91] J.-P. Champion, M. Loëte, and G. Pierre, *Spherical top spectra*, in Spectroscopy of the Earth's Atmosphere and Interstellar Medium, K. Narahari Rao and A. Weber, eds., Academic Press, Boston, 1992, pp. 339–422.

[92] J.-P. Champion and G. Pierre, *Vibration–rotation energies of harmonic and combination levels in tetrahedral $XY4$ molecules*, J. Mol. Spectrosc., 79 (1980), p. 255–280.

[93] B. I. Zhilinski í, V. I. Perevalov, and Vl. G. Tŷuterev, *Method of Irreducible Tensor Operators in the Theory of Molecular Spectra*, Nauka, Moscow, 1987. (In Russian.)

[94] J. Patera, R. T. Sharp, and P. Winternitz, *Polynomial irreducible tensors for point groups*, J. Math. Phys., 19 (1978), pp. 2362–2376.

[95] K. T. Hecht, *Vibration-rotation energies of tetrahedral XY_4 molecules* I. *Theory of spherical top molecules*, J. Mol. Spectrosc., 5 (1960), pp. 355–389.

[96] Ch. W. Patterson, *Quantum and semiclassical description of a triply degenerate anharmonic oscillator*, J. Chem. Phys., 83 (1985), pp. 4618–4632.

[97] K. Efstathiou, R. H. Cushman, and D. A. Sadovskií, *Linear Hamiltonian Hopf bifurcation for point-group-invariant perturbations of the* 1:1:1 *resonance*, R. Soc. London Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 2997–3019.

[98] K. Efstathiou and D. A. Sadovskií, *Perturbations of the* 1:1:1 *resonance with tetrahedral symmetry: A three degree of freedom analogue of the two degree of freedom Hénon–Heiles Hamiltonian*, Nonlinearity, 17 (2004), pp. 415–446.

[99] C. Jaffé and W. P. Reinhardt, *Time-independent methods in classical mechanics: Calculation of invariant tori and semiclassical energy levels via classical Van Vleck transformations*, J. Chem. Phys., 71 (1979), pp. 1862–1869.

[100] C. Jaffé and W. P. Reinhardt, *Uniform semiclassical quantization of regular and chaotic classical dynamics on the Henon–Heiles surface*, J. Chem. Phys., 77 (1982), pp. 5191–5203.

[101] A. B. McCoy and E. L. Sibert, *Calculation of infrared intensities of highly excited vibrational states of HCN using Van Vleck perturbation theory*, J. Chem. Phys., 95 (1991), pp. 3488–3493.

[102] A. B. McCoy and E. L. Sibert, *Perturbative calculations of vibrational* $(J = 0)$ *energy levels of linear molecules in normal coordinate representations*, J. Chem. Phys., 95 (1991), pp. 3476–3487.

[103] A. B. McCoy and E. L. Sibert, *An algebraic approach to calculating rotation-vibration spectra of polyatomic molecules*, Molecular Phys., 77 (1992), pp. 697–708.

[104] A. B. McCoy and E. L. Sibert, *Determining potential energy surfaces from spectra: An iterative approach*, J. Chem. Phys., 97 (1992), pp. 2938–2947.

[105] E. L. Sibert and A. B. McCoy, *Quantum, semiclassical and classical dynamics of the bending modes of acetylene*, J. Chem. Phys., 105 (1996), pp. 469–478.

[106] M. Joyeux, *Gustavson's procedure and the dynamics of highly excited vibrational states*, J. Chem. Phys., 109 (1998), pp. 2111–2122.

[107] D. Sugny and M. Joyeux, *On the application of canonical perturbation theory to floppy molecules*, J. Chem. Phys., 112 (2000), pp. 31–39.

[108] D. Sugny, M. Joyeux, and E. L. Sibert, *Investigation of the vibrational dynamics of the HCN/CNH isomers through high order perturbation theory*, J. Chem. Phys., 113 (2000), pp. 7165–7177.

[109] D. Sugny and M. Joyeux, *New canonical perturbation procedure for studying nonadiabatic dynamics*, Chem. Phys. Lett., 337 (2001), pp. 319–326.

[110] R. P. Stanley, *Invariants of finite groups and their applications to combinatorics*, Bull. Amer. Math. Soc., 1 (1979), pp. 475–511.

[111] B. Sturmfels, *Algorithms in Invariant Theory*, Springer-Verlag, Berlin, 1993.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 3, pp. 352–377

Tippe Top Inversion as a Dissipation-Induced Instability[*]

Nawaf M. Bou-Rabee[†], Jerrold E. Marsden[‡], and Louis A. Romero[§]

Abstract. By treating tippe top inversion as a dissipation-induced instability, we explain tippe top inversion through a system we call the modified Maxwell–Bloch equations. We revisit previous work done on this problem and follow Or's mathematical model [*SIAM J. Appl. Math.*, 54 (1994), pp. 597–609]. A linear analysis of the equations of motion reveals that the only equilibrium points correspond to the inverted and noninverted states of the tippe top and that the modified Maxwell–Bloch equations describe the linear/spectral stability of these equilibria. We supply explicit criteria for the spectral stability of these states. A nonlinear global analysis based on energetics yields explicit criteria for the existence of a heteroclinic connection between the noninverted and inverted states of the tippe top. This criteria for the existence of a heteroclinic connection turns out to agree with the criteria for spectral stability of the inverted and noninverted states. Throughout the work we support the analysis with numerical evidence and include simulations to illustrate the nonlinear dynamics of the tippe top.

Key words. tippe top inversion, dissipation-induced instability, constrained rotational motion, axisymmetric rigid body

AMS subject classifications. 70E18, 34D23, 37J15, 37M05

DOI. 10.1137/030601351

1. Introduction.

Tippe tops come in a variety of forms. The most common geometric form is a cylindrical stem attached to a truncated ball, as shown in Figure 1.1. On a flat surface, the tippe top will rest stably with its stem up. However, spun fast enough on its blunt end, the tippe top momentarily defies gravity, inverts, and spins on its stem until dissipation causes it to fall over. This spectacular sequence of events occurs because, and in spite of, dissipation.

Tippe top inversion is an excellent example of a dissipation-induced instability. Tippe top inversion is described by a system we call the modified Maxwell–Bloch equations. These

[*] Received by the editors October 22, 2003; accepted for publication (in revised form) by M. Golubitsky January 22, 2004; published electronically July 6, 2004. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siads/3-3/60135.html

[†] Applied and Computational Mathematics, Caltech, Pasadena, CA 91125 (nawaf@acm.caltech.edu). The research of this author was supported by the U.S. DOE Computational Science Graduate Fellowship through grant DE-FG02-97ER25308.

[‡] Control and Dynamical Systems, Caltech, Pasadena, CA 91125 (marsden@cds.caltech.edu). The research of this author was partially supported by the National Science Foundation.

[§] Sandia National Laboratories, P.O. Box 5800, MS 1110, Albuquerque, NM 87185-1110 (lromero@sandia.gov). The research of this author was supported by Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

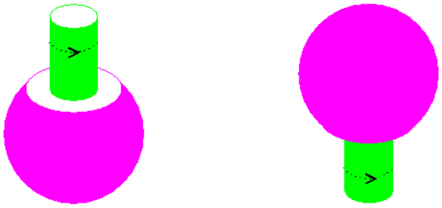


Figure 1.1. *From left: a sketch of the noninverted and inverted states of the tippe top.*

equations are a generalization of a previously derived normal form describing dissipation-induced instabilities in the neighborhood of the 1:1 resonance [4]. We will show in section 2 that the modified Maxwell–Bloch equations are the normal form for rotationally symmetric, planar dynamical systems.

A dissipation-induced instability describes a neutrally stable equilibrium becoming spectrally (and hence nonlinearly) unstable with the addition of dissipation. Dissipation-induced instability itself has a long history, which goes back to Thomson and Tait; see [14]. In its modern form, dissipation-induced instability was shown both to be a general phenomenon for gyroscopically stabilized systems and to provide a sharp converse to the energy momentum stability method by Bloch et al. in [1] and [2]. We refer the reader to these papers for additional examples and basic theory.

By considering tippe top inversion as a dissipation-induced instability, we observe precisely how tippe top inversion relates to well-understood dissipation-induced instabilities in gyroscopic systems.

History and literature. Tippe top inversion has been much investigated in the literature but never satisfactorily resolved. In what follows we survey a selection of theoretical results relevant to this paper. We refer the reader to the work of Cohen [5] and Or [11] for more comprehensive surveys of the literature.

Analysis of the tippe top dates back to the last century with Routh's commentary on the rising of tops. Routh's simple physical analysis made it clear that dissipation was fundamental in understanding the physics of tippe top inversion [12]. In 1977, Cohen confirmed Routh's physical analysis through numerical simulation [5]. Cohen modeled the tippe top as a holonomic ball with an inhomogeneous, axisymmetric mass distribution on a fixed plane. Thus, the ball's center of mass was on its axis of symmetry, but not coincident with its geometric center. Tippe top inversion corresponded to the ball's center of gravity moving above its geometric center. Cohen's mathematical model, derived using Newtonian mechanics, is written in terms of Euler angles and assumed a sliding friction law.

In reality, tippe top inversion is a transient phenomenon because of dissipation. However, by neglecting frictional torque, the spinning inverted state becomes a steady-state phenomenon. With a sliding friction law that neglects frictional torque, Cohen's equations of motion were amenable to a linear-stability analysis about the spinning inverted and noninverted states, but he did not carry out such an analysis in his work. In this sense the fact that the model neglects frictional torque is a feature rather than a defect of the friction law.

Cohen concluded that a sliding friction law based on Coulomb friction explains tippe top inversion. Moreover, Cohen showed that the inviscid ball (no sliding friction) does not invert. Coulomb friction is proportional to the normal force exerted by the surface at the point of contact and opposes the motion of the tippe top at the point of contact. This nonlinear friction law drops out of a standard linear-stability analysis.

In 1994, Or extended Cohen's work by generalizing his friction law, performing a dimensional analysis of the equations of motion, writing the equations of motion in convenient coordinates, and exploring the linearized behavior of the ball. To further facilitate a linear-stability analysis, Or added viscous friction to Cohen's friction law. The viscous friction is linearly related to the slip velocity at the point of contact.

Or's work showed that inversion could occur because of viscous or Coulomb friction. Or's analysis of tippe top inversion by Coulomb friction and viscous friction demonstrated viscous friction flips the top more rapidly than Coulomb friction. His analysis mentioned first integrals in the no-slip problem, but did not use them to explain the global behavior of the tippe top.

There was good reason to avoid a global analysis of the tippe top: the theoretical model of a ball did not accurately model the contact effects of the tippe top stem. However, the fundamental physics behind the global behavior of the tippe top was still reasonably described by the equations of motion for a ball with an eccentric center of mass, i.e., a gyroscopic, axisymmetric rigid body rising from a gravitationally favorable position to an unfavorable one because of dissipation. This observation motivated some aspects of the present study and possibly Ebenfeld's thesis on tippe top inversion from a Lyapunov perspective.

In 1995, Ebenfeld and Scheck applied a Lyapunov method to a dimensional version of Or's mathematical model. Using the energy as a Lyapunov function, they analyzed the orbital stability of the spinning, sliding ball. They showed that, starting with nonzero slip, the standing, spinning tippe top tends to manifolds of constant energy with no-slip and no tangential surface force. Moreover, they showed that without slip the tippe top cannot invert. They classified all possible asymptotic solutions as either tumbling (precession, spin, and no nutation) or rotating (pure spin) solutions. They provide criteria for determining the Lyapunov stability of these solutions [6].

We have applied some techniques of the present work to the related, but different, "rising egg" problem. (See Bou-Rabee, Marsden, and Romero [3].) In that paper we address the contributions of Moffatt [10] and Ruina [13] to that problem.

Main goals of this paper. The main result of this paper asserts that the stability of the noninverted and inverted tippe top (cf. Figure 1.1) and the existence of a heteroclinic connection between these states is completely described by the modified Maxwell–Bloch equations: a normal form for rotationally symmetric, planar dynamical systems. This result is important because (1) it shows how tippe top inversion relates to well-studied dissipation-induced instabilities in gyroscopic systems; (2) it integrates the linear and nonlinear analysis; and (3) it simplifies the analysis of tippe top inversion and yields explicit criteria for when the tippe top inverts.

We use Or's mathematical model with viscous friction to obtain equations of motion for the tippe top in a convenient set of coordinates. We locate and analyze the stability of all equilibrium points of the equations of motion. As Ebenfeld observed, the viscous problem conserves Υ_Q, the angular momentum about the vector connecting the center of mass of the ball to the surface point of contact. For a fixed value of Υ_Q, the only equilibrium points of the equations of motion correspond to the noninverted and inverted states of the tippe top. We extend Or's linear analysis by linearizing about both of these equilibrium points.

By assuming that the translation of the center of mass is negligible, we derive a reduced system for the tippe top in terms of only angular variables. We show that this reduced equation is of the form of a modified Maxwell–Bloch equation and give reasons to support this approximation. Analysis of explicit stability criteria for the reduced system shows that the reduced system accurately describes the linearized behavior of the tippe top. Thus, the paper shows that the stability of the noninverted and inverted tippe top is completely determined by the modified Maxwell–Bloch equations. Moreover, these equations are the normal form for tippe top inversion; i.e., they are the simplest possible equations that can capture tippe top inversion.

We extend the nonlinear analysis by proving the existence of a heteroclinic connection between the noninverted and inverted states of the tippe top. Application of LaSalle's invariance principle repeats/simplifies Ebenfeld's energy arguments. Like Ebenfeld, we use the energy as a Lyapunov function. For the viscous problem, we show that the energy's orbital derivative is negative semidefinite and a constant multiple of the ℓ_2-norm of the slip velocity. Thus, when the slip velocity vanishes, the energy is conserved. We identify the largest invariant set on this manifold characterized by no-slip and no tangential surface force.

The largest invariant set on this manifold extremizes the energy subject to the surface and angular momentum constraint: $\Upsilon_Q =$ constant. For certain parameter values, we find that the noninverted and inverted equilibrium states are the only extrema of the energy and are, in fact, the absolute maxima and minima of the energy, respectively. By the theorems of Barbashin and Krasovskii, trajectories starting in the neighborhood of the asymptotically unstable, noninverted state approach the asymptotically stable, inverted state as $t \to \infty$; i.e., the inverted state is globally asymptotically stable [8]. Outside this range of parameter values, the inverted and noninverted states become asymptotically unstable and a limit cycle corresponding to Ebenfeld's tumbling solution minimizes the energy.

Thus, the asymptotic states of the tippe top approach either (i) an isolated asymptotically stable equilibrium point or (ii) a neutrally stable limit cycle. The second case corresponds to the inverted and noninverted states being asymptotically unstable. In this case trajectories starting in the neighborhood of the noninverted, asymptotically unstable state will tend to limit cycles upon which total energy, magnitude of the angular momentum, nutation angle, and angular momentum about the vertical are conserved. We refer the reader to Ebenfeld for conditions for Lyapunov stability of this limit cycle. When the inverted state is asymptotically stable, a heteroclinic connection between the asymptotically unstable and stable states of the tippe top exists. We derive explicit criteria to determine the range of parameter values for which the heteroclinic connection exists.

A comparison of the existence criteria for the heteroclinic connection and the linear-stability criteria derived from the modified Maxwell–Bloch system for the tippe top shows they are explicitly related. Thus, the modified Maxwell–Bloch equations fully explain tippe top inversion.

Without dissipation, linear theory concludes that the equilibria are linearly (or neutrally) stable, which does not imply nonlinear stability. With dissipation, however, it is no surprise that the linear analysis provides the correct local nonlinear dynamics (since dissipation moves eigenvalues off of the imaginary axis).

Organization of the paper. In section 2, we present the modified Maxwell–Bloch equations. We derive these equations and supply specific stability criteria for the system's characteristic polynomial. We also discuss the ability of this model to capture the fundamental physics in tippe top inversion.

In section 3, we derive and discuss the mathematical model of the tippe top. Specifically, we write the dimensional equations of motion of the theoretical tippe top using Newtonian mechanics, discuss the friction law, and nondimensionalize these equations. We also explicitly compare the model in this paper to others in the literature.

In section 4, we apply linear theory to the nonlinear model. In particular, we locate the equilibria of the governing dimensionless equations, linearize about these states, as well as review and extend Or's numerical stability analysis.

In section 5, we cast the linearized equations for the tippe top in the form of the modified Maxwell–Bloch equations. We mention a direct derivation to these equations and analyze the stability of the modified Maxwell–Bloch equations. In particular, we show that these equations are the simplest possible equations that can capture the fundamental physics in tippe top inversion.

In section 6, we explain tippe top inversion from an energy landscape perspective. We apply LaSalle's invariance principle to determine the asymptotic state of the tippe top, study the behavior of solutions on this asymptotic state, and determine when this asymptotic state corresponds to the inverted tippe top. We also compare the criteria for the existence of a heteroclinic connection with the linear-stability criteria for the tippe top modified Maxwell–Bloch equations provided in section 5.

In section 7, we discuss data from simulations which verify the local and global analysis in the paper. We conclude the paper with a discussion of future directions inspired by this work.

2. Modified Maxwell–Bloch equations.

This section introduces an important extension of the Maxwell–Bloch equations and studies their stability.

Derivation. Consider a planar ODE of the form

$$\ddot{q} = f(q, \dot{q}), \qquad q = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Linearization of these equations yields

$$\ddot{q} = A\dot{q} + Bq.$$

The characteristic polynomial of this system

$$\det\left(\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} + A\sigma + B \right) = 0$$

shows that the ODE will have time-reversal symmetry; i.e., if σ is a solution, then so is $-\sigma$, when A is skew-symmetric and B is symmetric.

We define the rotation matrix

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

as well as the identity and elementary skew-symmetric matrix in $\mathbb{R}^2$:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad S = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The necessary and sufficient condition for a 2×2 matrix in $\mathbb{R}^2$ to commute with the rotation matrix is that the matrix be a linear combination of I and S.

Thus, if this ODE is rotationally symmetric, i.e., the ODE is invariant under SO(2) rotation, then the matrices A and B can be expressed as

$$A = -\alpha S - \beta I, \qquad B = -\gamma S - \delta I,$$

where α, β, γ, and δ are real scalars. Because α and γ can destroy time-reversal symmetry in A and B, we call these terms dissipative.

Given the particular form of the rotationally symmetric ODE, we can write the two-dimensional real system as a one-dimensional complex system,

$$\ddot{z} + i\alpha\dot{z} + \beta\dot{z} + i\gamma z + \delta z = 0, \qquad z = x + iy, \tag{2.1}$$

which we call the *modified Maxwell–Bloch equations*. We observe that (2.1) is the basic harmonic oscillator with the two complex terms $i\alpha\dot{z}$ and $i\gamma z$. In physical systems, the first term arises from Coriolis effects, and hence is known as the gyroscopic term. The second term typically arises from dissipation in rotational variables. This damping force is different from the usual damping term proportional to absolute velocity, $\beta\dot{z}$. Physically the complex damping term models viscous effects caused by, for example, motion in a fluid, while the usual damping term models internal dissipation.

Proposition 2.1. *The modified Maxwell–Bloch equations are the linearized normal form for planar, rotationally symmetric dynamical systems.*

Stability criteria. The characteristic polynomial of the modified Maxwell–Bloch equations is

$$\lambda^4 + 2\beta\lambda^3 + (\alpha^2 + \beta^2 + 2\delta)\lambda^2 + 2(\alpha\gamma + \beta\delta)\lambda + (\gamma^2 + \delta^2) = 0.$$

We now write the necessary and sufficient conditions for this polynomial to be Hurwitz [7].

Proposition 2.2. *The zero solution of the modified Maxwell–Bloch equations is asymptotically stable provided that the following inequalities hold:*

$$\begin{aligned} \beta &> 0, \\ \alpha\beta\gamma - \gamma^2 + \beta^2\delta &> 0, \\ \alpha^2\beta + \beta^3 - \alpha\gamma + \beta\delta &> 0. \end{aligned} \tag{2.2}$$

There are two especially interesting physical cases of these equations:

1. When $\delta > 0$, $\gamma = \beta = 0$, the system is neutrally stable with or without the presence of the gyroscopic term. Adding usual dissipation ($\beta > 0$) makes the neutrally stable zero solution asymptotically stable. Adding, however, damping in rotational variables can stabilize or destabilize the neutrally stable zero solution.
2. When $\delta < 0$, $\alpha > -4\delta > 0$, $\beta = \gamma = 0$, the system is gyroscopically, and hence neutrally, stable. Adding usual damping makes the neutrally stable zero solution asymptotically unstable since the second inequality in (2.2) can never hold. This case corresponds to the classical dissipation-induced instability [1]. If $\beta = 0$ and $\beta > 0$, the neutrally stable zero solution becomes asymptotically unstable. Adding damping in both variables, i.e., $\beta > 0$ and $\gamma > 0$, can stabilize or destabilize the zero solution depending on the ratio of β to γ.

For the tippe top, we will show that dissipation in rotational variables (or complex damping) is essential to understanding inversion. In fact, the remarks above point out some limitations of usual damping: usual damping can only predict instability in the case of a gyroscopically stable system and stability in the case of a gravitationally stable system.

Consider the modified Maxwell–Bloch equations as a possible model of the linearized behavior of the tippe top. In particular, suppose that the noninverted and inverted states of the tippe top correspond to the zero solution of (2.1). In the inviscid case, we observe a noninverted state which is gravitationally stable with or without gyroscopic effects. Remark 1 above shows that the addition of usual damping cannot destabilize this gravitationally stable, noninverted state. The complex damping term, however, can destabilize this state. Therefore, the complex damping term can explain why the gravitationally stable tippe top becomes asymptotically unstable.

Moreover, after the tippe top inverts we have a gyroscopically stabilized inverted state. We have shown that the addition of usual damping would make such a system asymptotically unstable. Thus, usual damping cannot explain why the tippe top spins stably in its inverted state. Remark 2 shows that the complex and usual damping term in the right ratio can, however, stabilize this state. Thus, the complex damping term can also explain why the tippe top spins stably on its stem. We will revisit this analysis when we cast the linearized equations of the tippe top in the form of the modified Maxwell–Bloch equations.

3. Tippe top governing equations.

This section contains a pedagogical derivation of the tippe top governing equations following Or [11], given mainly for the reader's convenience. The section concludes with a brief discussion of how to explicitly obtain the equations derived by Or [11] and Ebenfeld and Scheck [6] from our form of the equations and how to formulate the equations of motion using Lagrangian mechanics.

Derivation. We write the equations of motion of the tippe top from first principles. We idealize the tippe top as a ball of radius R and mass M on a fixed plane. The mass distribution of the ball is inhomogeneous, but symmetric about an axis through the ball's geometric center. Thus, the ball's center of mass is located on the axis of symmetry $\mathbf{k}$, which is pointing in the direction (l, m, n), but at a distance $Re^\star$ above the geometric center, where $e^\star$ is the center of mass offset $(0 \le e^\star \le 1)$.

Let the points Q, O, and C represent the point of contact, the geometric center, and the center of mass of the ball, respectively (cf. Figure 3.1). We define $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ to be unit vectors of a nonrotating Cartesian frame attached to O. We define principal axes as $\mathbf{i}, \mathbf{j}, \mathbf{k}$ with axis of symmetry $\mathbf{k}$. Let $\mathbf{L}, \boldsymbol{\omega}, (X, Y, Z)$ be the ball's angular momentum about O, angular velocity, and absolute position of its center of mass, respectively. We define I_k and $I = I_i = I_j$ to be the inertias about the respective principal axes attached to O.

Let the vector $\mathbf{Q}$ represent the position of the point of contact Q with respect to the center of mass O given by

$$\mathbf{Q} = R(-e^\star \mathbf{k} - \mathbf{e}_z). \tag{3.1}$$

We assume that the only external forces acting on the tippe top are due to gravity and the surface force $\mathbf{F}_Q$ at the point of contact Q. Conservation of linear momentum yields

$$\begin{aligned} M\ddot{X} &= \mathbf{F}_Q \cdot \mathbf{e}_x, \\ M\ddot{Y} &= \mathbf{F}_Q \cdot \mathbf{e}_y, \\ M\ddot{Z} &= \mathbf{F}_Q \cdot \mathbf{e}_z - Mg. \end{aligned} \tag{3.2}$$

By the translation theorem of angular momentum, we have

$$\dot{\mathbf{L}} = MR^2(e^\star)^2 \mathbf{k} \times \ddot{\mathbf{k}} + \mathbf{Q} \times \mathbf{F}_Q. \tag{3.3}$$

Since the axis of symmetry undergoes pure rotation, we have

$$\dot{\mathbf{k}} = \boldsymbol{\omega} \times \mathbf{k},$$

which is known as the attitude equation. By projecting the angular momentum on the principal coordinate frame, we can relate $\boldsymbol{\omega}$ and $\mathbf{L}$ as follows:

$$\mathbf{L} = I\boldsymbol{\omega} + \frac{I_k - I}{I_k}(\mathbf{L} \cdot \mathbf{k})\mathbf{k}.$$

Solving for $\boldsymbol{\omega}$ we obtain

$$\boldsymbol{\omega} = \frac{1}{I}\left(\mathbf{L} + \frac{I_k - I}{I_k}(\mathbf{L} \cdot \mathbf{k})\mathbf{k}\right).$$

Substituting this expression into the attitude equation, we obtain

$$\dot{\mathbf{k}} = \frac{1}{I}\mathbf{L} \times \mathbf{k}. \tag{3.4}$$

We also have the constraint that keeps the ball on the fixed plane (no hopping or penetration into the surface):

$$\mathbf{Q} \cdot \mathbf{e}_z + z = 0. \tag{3.5}$$

Together, (3.2), (3.3), (3.4), and (3.5) represent the dimensional equations of motion of the ball once the force $\mathbf{F}_Q$ has been specified.

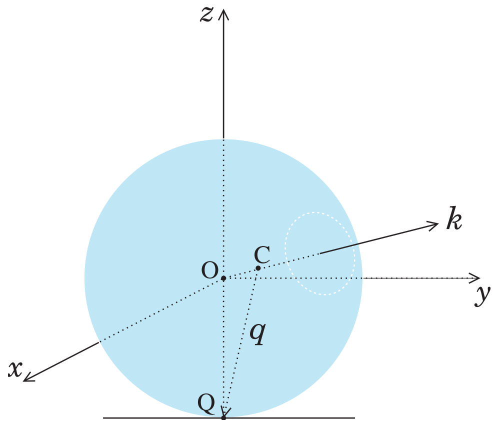


Figure 3.1. *We model the tippe top as a ball with an eccentric center of mass C, geometric center O, and point of contact Q. Vectors $\mathbf{q}$ and $\mathbf{k}$ represent the dimensionless position of the contact point with respect to the center of mass and the unit vector in the direction of the axis of symmetry, respectively.*

Friction law. The force exerted on the body at the contact point Q is due to surface frictional and normal reaction forces: $\mathbf{F}_Q = \mathbf{F}_f + F_z\mathbf{e}_z$. It is well understood that sliding friction is the main mechanism behind tippe top inversion. To facilitate the linear and nonlinear analysis, we assume a sliding frictional force proportional to the slip velocity, i.e., the velocity of the contact point on the rigid body relative to the center of mass $\mathbf{V}_Q$:

$$\mathbf{F}_f = -c\mathbf{V}_Q.$$

A more complete friction law would include rotational, pure rolling, and Coulomb friction. We neglect frictional torque so that the spinning, standing ball is a steady-state solution of the equations of motion. Addition of nonlinear Coulomb friction would result in algebraic rather than exponential divergence of the unstable equilibrium solution [11]. Pure rolling friction would be important in the limit to a perfectly rough surface.

The slip velocity, i.e., the velocity of the contact point on the rigid body relative to the center of mass, is

$$\mathbf{V}_Q = \mathbf{V}_C + \boldsymbol{\omega} \times \mathbf{Q}, \tag{3.6}$$

where $\mathbf{V}_C = (\dot{X}, \dot{Y}, \dot{Z})$ is the absolute velocity of the center of mass. The slip velocity in terms of the angular momentum is given by

$$\mathbf{V}_Q = \mathbf{V}_C + \frac{R}{I}\left((e^\star\mathbf{k} + \mathbf{e}_z) \times \mathbf{L} + \frac{I_k - I}{I_k}(\mathbf{L}\cdot\mathbf{k})(\mathbf{e}_z \times \mathbf{k})\right).$$

Dimensionless equations. We introduce the following dimensionless variables:

$$x = \frac{X}{R}, \quad y = \frac{Y}{R}, \quad z = \frac{Z}{R}, \quad t = T\Omega, \quad \mathbf{f}_Q = \frac{R\mathbf{F}_Q}{I_k\Omega^2}, \quad \boldsymbol{\Upsilon} = \frac{\mathbf{L}}{I_k\Omega}, \quad \mathbf{q} = \frac{\mathbf{Q}}{R},$$

and parameters

$$\sigma = \frac{I_k}{I}, \quad \mathrm{Fr}^{-1} = \frac{g}{\Omega^2 R}, \quad \mu = \frac{MR^2}{I}, \quad \nu = \frac{cR^2}{I\Omega},$$

where Ω is the spin rate of the initially standing equilibrium solution we will linearize about and T represents dimensional time. The dimensionless parameters σ, Fr, μ, and ν are the inertia ratio, Froude number, dimensionless mass, and friction factor, respectively.

The governing equations follow:

$$\begin{aligned} \mu\ddot{x} &= \sigma f_x = -\nu[\dot{x} - e^\star \dot{l} - \sigma \Upsilon_y + (\sigma - 1)(\boldsymbol{\Upsilon}\cdot\mathbf{k})m], \\ \mu\ddot{y} &= \sigma f_y = -\nu[\dot{y} - e^\star \dot{m} + \sigma \Upsilon_x - (\sigma - 1)(\boldsymbol{\Upsilon}\cdot\mathbf{k})l], \\ \dot{\boldsymbol{\Upsilon}} &= \frac{\left[\sigma^2\mu(e^\star)^2(\boldsymbol{\Upsilon}\cdot\mathbf{k})(\mathbf{k}\times\boldsymbol{\Upsilon}) + e^\star\sigma f_z(\mathbf{e}_z\times\mathbf{k}) + \sigma\mathbf{q}\times\mathbf{f}_f - \sigma\mu(e^\star)^2(\mathbf{f}_f\cdot(\mathbf{k}\times\mathbf{e}_z))\mathbf{k}\right]}{(1-\mu(e^\star)^2)\sigma}, \\ \dot{\mathbf{k}} &= \sigma\boldsymbol{\Upsilon}\times\mathbf{k}. \end{aligned} \tag{3.7}$$

Equation (3.7) describes the translational (x, y) and rotational $(\boldsymbol{\Upsilon}, \mathbf{k})$ motion of the ball.

To determine the normal reaction force f_z we use the constraint that keeps the ball on the surface,

$$z = 1 + e^\star(\mathbf{e}_z\cdot\mathbf{k}), \tag{3.8}$$

together with the vertical translation equation,

$$\ddot{z} = \frac{\sigma}{\mu} f_z - \mathrm{Fr}^{-1}. \tag{3.9}$$

The appendix gives an explicit expression for the normal reaction force.

Observe that the angular momentum about the vector connecting the center of mass to the contact point $\mathbf{q}$ is a conserved quantity for (3.7):

$$\Upsilon_Q = \Upsilon_{cg}\cdot\mathbf{q} = -\sigma e^\star(\boldsymbol{\Upsilon}\cdot\mathbf{k})(1 + \sigma\mu e^\star n) - \sigma(1 - \sigma\mu(e^\star)^2)\Upsilon_z, \qquad (\Upsilon_Q)_t = 0. \tag{3.10}$$

This first integral, known in the literature as the Jellett invariant, can be derived geometrically from a momentum map. The Jellett invariant corresponds to an $\mathbb{S}^1$ action on the configuration space $Q = \mathrm{SO}(3)\times\mathbb{R}^2$, which can be computed by formula (12.2.1) of [9]. Specifically, the action for $\theta\in\mathbb{S}^1$ is given by simultaneous rotation about the axis of symmetry by the angle $-\sigma e^\star(1+\sigma\mu e^\star n)\theta$ and about the vertical by the angle $(1-\sigma\mu(e^\star)^2)\theta$. Because the force at the point of contact does no virtual work under this action, dissipation does not destroy this conservation law.

We briefly compare this model with others in the literature. Writing the absolute velocity of the center of mass in terms of the absolute velocity of the center of curvature leads to the form of the equations that can be found in Or's work [11, eqs. (1) and (2), p. 600]:

$$\begin{aligned} \mathbf{V}_0 &= \mathbf{v}_C - e^\star\sigma\boldsymbol{\Upsilon}\times\mathbf{k}, \\ \boldsymbol{\omega} &= \frac{1}{\sigma}(\boldsymbol{\omega} + (\sigma-1)(\boldsymbol{\omega}\cdot\mathbf{k})\mathbf{k}). \end{aligned}$$

We choose to write the governing equations in terms of the velocity of the center of mass because we are interested in the translational effects of the ball's center of mass. Writing the angular momentum with respect to the center of mass C leads to the form of the equations found in Ebenfeld's work [6, eq. (14), p. 201]:

$$\Upsilon_{cg} = (1 - \sigma\mu(e^\star)^2)\boldsymbol{\Upsilon} + \sigma\mu(e^\star)^2(\boldsymbol{\Upsilon}\cdot\mathbf{k})\mathbf{k}.$$

In accordance with Or's equations, we choose to write the angular momentum with respect to the center of curvature O.

The basic phase space is TQ, which is typically parametrized by rotational and translational coordinates and their conjugate momenta. The body angular velocity is $\mathbf{R}^{-1}\dot{\mathbf{R}}$, where $\mathbf{R}$ is the attitude matrix of the body and $(\mathbf{R}, \dot{\mathbf{R}})$, together with translational coordinates, is a point in TQ. The body angular momentum coordinates are then related to the body angular velocity coordinates by constant factors given by the principal moments of inertia. The Lagrangian $\mathcal{L} : TQ \to \mathbb{R}$ in terms of these variables is given explicitly by

$$\mathcal{L} = \mu\left(\frac{\dot{x}^2}{2} + \frac{\dot{y}^2}{2} + \frac{\dot{z}^2}{2}\right) + \sigma(1 - \sigma\mu(e^\star)^2)\frac{\boldsymbol{\Upsilon}\cdot\Upsilon}{2} + \frac{(1 - \sigma + \sigma\mu(e^\star)^2)(\boldsymbol{\Upsilon}\cdot\mathbf{k})^2}{2} - \mu\mathrm{Fr}^{-1}z.$$

The equations of motion can then be formulated by the Lagrange d'Alembert principle with generalized forces given by the specified sliding friction force. The surface reaction force is a force of constraint that is an intrinsic part of the Euler–Lagrange operator.

4. Linear theory.

Equilibria. Equilibria of (3.7) satisfy

$$\begin{aligned} \boldsymbol{\Upsilon}\times\mathbf{k} = 0 &\implies \boldsymbol{\Upsilon} \text{ and } \mathbf{k} \text{ are collinear,}\\ \mathbf{e}_z\times\mathbf{k} = 0 &\implies \mathbf{e}_z \text{ and } \mathbf{k} \text{ are collinear.} \end{aligned}$$

Therefore, equilibria satisfy

$$\dot{x} = \dot{y} = x = y = \Upsilon_x = \Upsilon_y = l = m = 0, \quad \Upsilon_z = \text{constant}, \quad n = \pm 1.$$

If we restrict the angular momentum about $\mathbf{q}$ (cf. (3.10)) to a certain value, i.e., $\Upsilon_Q = -(1 + e^\star)\sigma$, we have two equilibria given by

$$\dot{x} = \dot{y} = x = y = \Upsilon_x = \Upsilon_y = l = m = 0, \quad \Upsilon_z = 1, \frac{1 + e^\star}{1 - e^\star}, \quad n = \pm 1. \tag{4.1}$$

These equilibria correspond to the noninverted and inverted states of the tippe top shown in Figure 1.1.

Linearization. Linearizing the equations of motion about these states ($n = n^o = \pm 1$, $\Upsilon_z = \Upsilon_z^o$) and writing the resulting equations in terms of the complex variables

$$V_C = \dot{x} + i\dot{y}, \quad \Lambda = \Upsilon_x + i\Upsilon_y, \quad \Phi = l + im,$$

we obtain

$$\begin{aligned}\dot{V}_C &= -\nu A V_C - i\nu B\Lambda + i\nu C\Phi,\\ \dot{\Lambda} &= i\nu D V_C + (\nu E + iF)\Lambda + (\nu G + iH)\Phi,\\ \dot{\Phi} &= -in^o\sigma\Lambda + i\sigma\Upsilon_z^o\Phi,\end{aligned} \tag{4.2}$$

where we have introduced the following real parameters to ensure clarity of (4.2):

$$A = \frac{1}{\mu}, \quad B = \frac{1}{\mu}\sigma(e^\star n^o + 1), \quad C = \frac{(\sigma e^\star + n^o(\sigma - 1))\Upsilon_z{}^o}{\mu}, \quad D = \frac{(e^\star n^o + 1)}{\sigma(1 - \mu(e^\star)^2)},$$

$$E = \frac{(e^\star n^o + 1)^2}{1 - \mu(e^\star)^2}, \quad F = \frac{n^o\mu(e^\star)^2\sigma}{1 - \mu(e^\star)^2},$$

$$G = \frac{(e^\star n^o + 1)(\sigma e^\star + n^o(\sigma - 1))\Upsilon_z{}^o}{\sigma(1 - \mu(e^\star)^2)}, \quad H = \frac{-\mu(e^\star)^2\sigma^2\Upsilon_z{}^o + \mu e^\star \mathrm{Fr}^{-1}}{\sigma(1 - \mu(e^\star)^2)}.$$

To obtain Or's linearized model, set $n^o = 1$, $\Upsilon_z{}^o = 1$ in (4.2) and express the equations in terms of the velocity of the center of curvature, V_0, and the angular velocity, $\boldsymbol{\omega}$, through the following change of dependent variables: $V_C = V_0 + e^\star i(\Phi - \boldsymbol{\omega})$, and $\sigma\Lambda = \boldsymbol{\omega} + (\sigma - 1)\Phi$ [11, eq. (11), p. 602].

If we assume a solution of the form $e^{\lambda t}\phi$, we obtain the following eigenvalue problem:

$$\lambda\phi = \nu\mathbf{C}\phi + i\mathbf{K}\phi, \tag{4.3}$$

where

$$\mathbf{C} = \begin{bmatrix} -A & -iB & iC \\ iD & E & G \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & iF & iH \\ 0 & -n^o\sigma & \Upsilon_z{}^o\sigma \end{bmatrix}.$$

Numerical stability analysis. Here we briefly confirm Or's numerical stability results and discuss the stability of the inverted state: $n^o = -1, \Upsilon_z{}^o = \frac{1+e^\star}{1-e^\star}$. A numerical linear stability analysis of the governing equation reveals the effect of the Froude number Fr^{-1}, the inertia ratio σ, and the friction factor ν as a function of the eccentricity $e^\star$. Figure 4.1 shows the growth factor of the eigenvalue with largest real part. When the center of mass is below the center of curvature, we see that a high Froude number destabilizes the noninverted state. A low Froude number implies that gravitational effects are more important than rotational effects, and therefore if the center of mass is below the center of curvature, a low Froude number stabilizes the standing spinning solution. When the center of mass moves above the center of curvature, we see that the Froude number has the opposite effect.

Figure 4.3 shows that the inertia ratio σ destabilizes the standing spinning solution for $\sigma < 1.14$ when the center of mass is below the center of curvature. As Cohen notes, the inertia ratio of most commercial tippe tops is less than 1, i.e., $\sigma < 1$.

Figure 4.5 shows that the inviscid problem is neutrally stable, as expected. As ν increases, the magnitude of the real part of the eigenvalue initially becomes larger until the friction

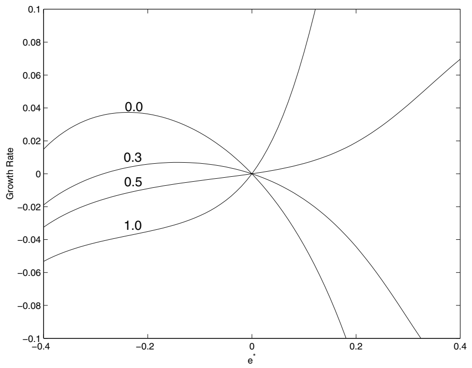


Figure 4.1. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for varying Froude numbers ($Fr^{-1} = 0, 0.3, 0.5, 1.0$) as functions of $e^\star$ for the noninverted equilibrium state $n^o = 1$, $\Upsilon_z{}^o = 1$.*

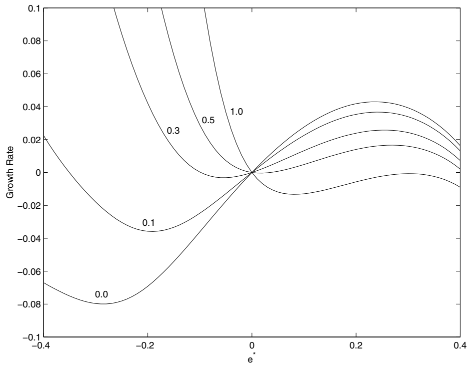


Figure 4.2. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for varying Froude numbers ($Fr^{-1} = 0, 0.3, 0.5, 1.0$) as functions of $e^\star$ for the inverted equilibrium state $n^o = -1$, $\Upsilon_z{}^o = \frac{1+e^\star}{1-e^\star}$.*

becomes so large that the ball ceases to slip. The real part of the eigenvalue then begins to decay as the ball tends to roll without slip. In the limit to pure rolling, we see that the nonholonomic problem is also neutrally stable. For $\nu > 0$, we also observe that the real part of the eigenvalue is positive (or negative) when the center of gravity of the ball is below (or above) the geometric center. This explains why the inverted state is linearly stable. All of these findings agree with those of Or [11].

Figures 4.2, 4.4, and 4.6 show that the behavior of the inverted state when $e^\star < 0$ is qualitatively similar to the behavior of the noninverted state with $e^\star > 0$; however, they are not quantitatively the same. In all of these figures, we see ranges of parameter values where the inverted state is asymptotically stable and unstable when $e^\star < 0$. We will show later that for parameter values where the inverted state is asymptotically stable and the noninverted state is asymptotically unstable, a heteroclinic orbit exists, and the inverted state is globally asymptotically stable.

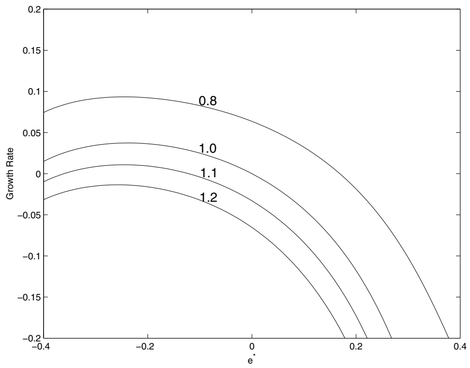


Figure 4.3. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for varying inertia ratios ($\sigma = 0.8, 1.0, 1.1, 1.2$) as functions of $e^\star$ for the noninverted equilibrium state $n^o = 1$, ${\Upsilon_z}^o = 1$.*

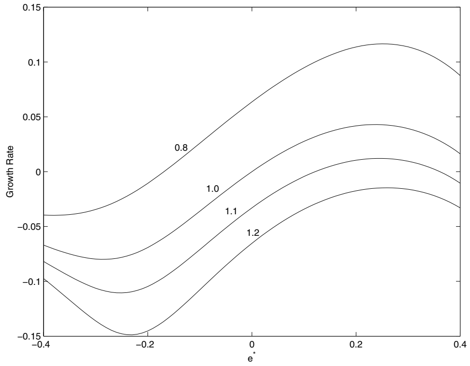


Figure 4.4. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for varying inertia ratios ($\sigma = 0.8, 1.0, 1.1, 1.2$) as functions of $e^\star$ for the inverted equilibrium state $n^o = -1$, ${\Upsilon_z}^o = \frac{1+e^\star}{1-e^\star}$.*

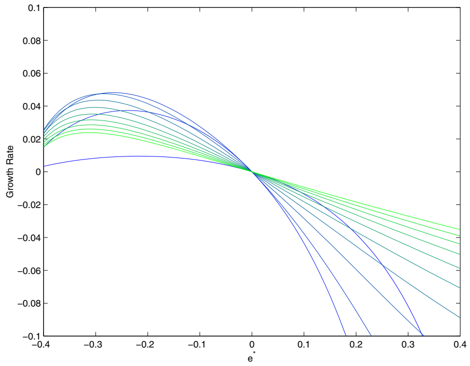


Figure 4.5. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for $Fr^{-1} = 0$ as functions of $e^\star$ for ν varying from 0.1 (blue) to 5.0 (green) for the noninverted equilibrium state $n^o = 1$, ${\Upsilon_z}^o = 1$.*

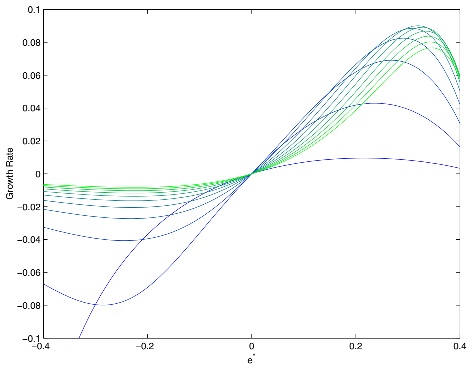


Figure 4.6. *The growth factor of the eigenvalue of* (4.3) *with largest real part is plotted here for* $Fr^{-1} = 0$ *as functions of* $e^\star$ *for* ν *varying from* 0.1 *(blue) to* 5.0 *(green) for the inverted equilibrium state* $n^o = -1$, $\Upsilon_z{}^o = \frac{1+e^\star}{1-e^\star}$.

5. Tippe top modified Maxwell–Bloch equations.

If we ignore translational effects, the linearized equations (4.2) simplify:

$$\begin{aligned}\dot{\Lambda} &= (\nu E + iF)\Lambda + (\nu G + iH)\Phi,\\ \dot{\Phi} &= -in^o\sigma\Lambda + i\sigma\Upsilon_z{}^o\Phi.\end{aligned} \tag{5.1}$$

To obtain these equations more directly, we can ignore translational effects from the outset. The derivation based on this assumption is closely related to deriving the equations of motion for the standard top in a gravitational field. There is ample evidence supporting this assumption:

1. The assumption is rigorously true for small friction. Equation (4.2) shows that when the friction is zero the velocity of the center of mass is fixed. Now consider a perturbation of an inviscid, initially restating state in terms of the dissipation ν. To first order in ν, it is sufficient to use only the zeroth-order velocity of the center of mass, which is by assumption zero. Thus, to first order the center of mass can be assumed to remain fixed. This reasoning can be borne out by application of rigorous perturbation theory. We do not attempt such an analysis in this work.
2. Numerical evidence shows that application of the perturbation theory of eigenvalues with or without the effect of translation produces the same stability results.
3. We will show that the position of the center of mass remains fixed for all extrema of the constrained energy (see section 6). Thus, translational effects are negligible in the steady-state behavior of trajectories in the neighborhood of equilibria (critical points of the constrained energy).
4. Simulation shows that the translational energy of the tippe top is negligible (see Figure 7.2) and that the tippe top rotates much more than it translates.

Equation (5.1) can be rewritten in terms of Φ alone:

$$\ddot{\Phi} + ia\dot{\Phi} + b\dot{\Phi} + ic\Phi + d\Phi = 0, \tag{5.2}$$

where

$$\begin{aligned}
a &= \frac{\sigma(\Upsilon_z{}^o - (\Upsilon_z{}^o - n^o)(e^\star)^2\mu)}{-1+(e^\star)^2\mu}, \\
b &= -\frac{(1+n^o e^\star)^2\nu}{-1+(e^\star)^2\mu}, \\
c &= \frac{\nu\Upsilon_z{}^o(1+n^o e^\star)(-(n^o)^2(-1+\sigma)+\sigma)}{-1+(e^\star)^2\mu}, \\
d &= \frac{\mathrm{Fr}^{-1}e^\star\mu n^o}{-1+(e^\star)^2\mu}.
\end{aligned}$$

Notice that (5.2) is in the form of the modified Maxwell–Bloch equations (2.1).

Now let us check the stability criteria for modified Maxwell–Bloch systems given in (2.2). The stability criteria for the noninverted state ($n^o = 1$, $\Upsilon_z{}^o = 1$) are

$$\begin{aligned}
1-\mu(e^\star)^2 &> 0, \\
(1+e^\star)\nu(\mathrm{Fr}^{-1}e^\star(1+e^\star)\mu(-1+\mu(e^\star)^2)+(1+e^\star)^5\nu^2+\sigma(-1+\sigma+e^\star(e^\star\mu+\sigma))) &> 0, \\
-(1+e^\star)^2\nu^2(1+\mathrm{Fr}^{-1}e^\star(1+e^\star)^2\mu-\sigma-e^\star(e^\star\mu+\sigma)) &> 0.
\end{aligned} \tag{5.3}$$

Likewise, the stability criteria for the inverted state ($n^o = -1$, $\Upsilon_z{}^o = \frac{1+e^\star}{1-e^\star}$) are

$$\begin{aligned}
&1-\mu(e^\star)^2 > 0, \\
&\nu\frac{(-\mathrm{Fr}^{-1}(-1+e^\star)^3e^\star\mu(-1+\mu(e^\star)^2)+(-1+e^\star)^7\nu^2)}{(1-e^\star)(-1+\mu(e^\star)^2)} \\
&+\nu\frac{(-1-e^\star+2(e^\star)^2\mu)\sigma(-1-e^\star+\sigma+(e^\star)^2(\mu-\sigma-2\mu\sigma)+(e^\star)^3(\mu+2\mu\sigma))}{(1-e^\star)(-1+\mu(e^\star)^2)} > 0, \\
&\nu^2(-1+\sigma+e^\star(-2+\mathrm{Fr}^{-1}\mu+\sigma)) \\
&+\nu^2((e^\star)^2(-1+(1+e^\star)^2+\mathrm{Fr}^{-1}(-2+e^\star)(2+(-2+e^\star)e^\star))\mu-\sigma-e^\star\sigma) > 0.
\end{aligned} \tag{5.4}$$

For the parameter values explored, the first two inequalities are always satisfied. The last inequality is satisfied precisely in the regions where the growth rate of the eigenvalue of (4.3) with largest real part is negative. We observe that this inequality is independent of the friction factor unless $\nu = 0$. Thus, we conclude that stability is independent of the magnitude of ν. Figures 5.1 and 5.2 superpose plots of the growth rate with the value of the last inequality in (5.3).

Can we reduce (5.2) any further? The answer is no because of the remarks made in section 2. In particular, it can be shown that, without the complex and usual damping terms, i.e., $b = 0$ and $c = 0$ or $\nu = 0$, the gravitationally stable noninverted state cannot become asymptotically unstable. Moreover, the gyroscopically stabilized state can be asymptotically stable if and only if the complex and usual damping terms are present and in the right ratio. In summary, the normal form for the tippe top in the neighborhood of inversion (a dissipation-induced instability) is described by the modified Maxwell–Bloch equations.

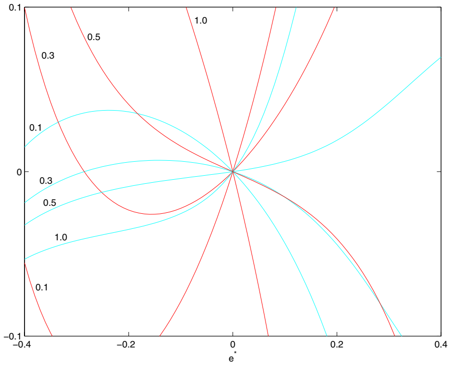


Figure 5.1. *The growth factor of the eigenvalue of* (4.3) *with largest real part (cyan curves) and the third inequality in the stability criteria given in* (5.3) *(red curves) are plotted here for varying Froude numbers ($Fr^{-1} = 0, 0.3, 0.5, 1.0$) as functions of $e^\star$. Here we consider the noninverted state $n^o = 1$, ${\Upsilon_z}^o = 1$.*

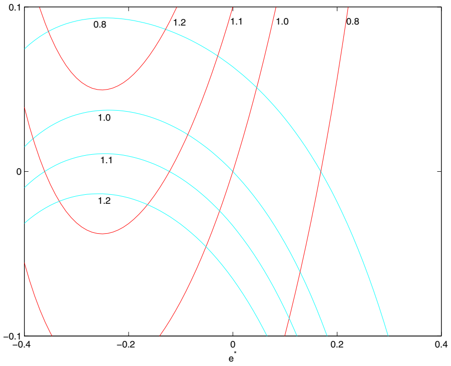


Figure 5.2. *The growth factor of the eigenvalue of* (4.3) *with largest real part (cyan curves) and the third inequality in the stability criteria given in* (5.3) *(red curves) are plotted here for varying inertia ratios ($\sigma = 0.8, 1.0, 1.1, 1.2$) as functions of $e^\star$. Here we consider the noninverted state $n^o = 1$, ${\Upsilon_z}^o = 1$.*

6. Heteroclinic connection.

Tippe top asymptotic states. To establish the existence of a heteroclinic orbit connecting the asymptotically stable and unstable states of the theoretical tippe top, we will invoke LaSalle's invariance principle. Consider a vector field χ on a manifold P. Let V be a Lyapunov function with negative semidefinite orbital derivative: $V_t \leq 0$ for all $z \in P$. We define the set $\aleph := \{z \in P | V_t(z) = 0\}$.

Theorem 6.1 (LaSalle's principle). *Let $z : [0, \infty) \to P$ be an integral curve of a vector field χ with initial condition $z(0) = z_0$. Suppose there is a positively invariant set (trapping region) M such that $z(t) \in M$ for all $t \geq 0$. Then $z(t)$ converges to the largest subset of $\aleph \cap M$ that is invariant under the flow of χ for all t, positive and negative.*

We will now apply this principle to (3.7), the governing equation of the viscous ball. We write the energy of the viscous ball as

$$E = \mu\left(\frac{\dot{x}^2}{2} + \frac{\dot{y}^2}{2} + \frac{\dot{z}^2}{2}\right) + \sigma(1 - \sigma\mu(e^\star)^2)\frac{\boldsymbol{\Upsilon}\cdot\boldsymbol{\Upsilon}}{2} + \frac{(1 - \sigma + \sigma\mu(e^\star)^2)(\boldsymbol{\Upsilon}\cdot\mathbf{k})^2}{2} + \mu\mathrm{Fr}^{-1}z, \tag{6.1}$$

which is a sum of translational, rotational, and gravitational effects. The energy's orbital derivative is given by

$$E_t = -\nu \|\mathbf{v}_Q\|^2, \tag{6.2}$$

where $\|\mathbf{v}_Q\|$ is the ℓ_2-norm of the dimensionless slip velocity. The energy's orbital derivative implies that the energy decreases monotonically until the slip velocity vanishes.

No-slip no-force problem. We now mention some properties of the asymptotic state, where the slip velocity vanishes and energy is conserved. The governing equations (3.7) without slip reduce to the system

$$\begin{aligned} \ddot{x} &= 0, \\ \ddot{y} &= 0, \\ \dot{\boldsymbol{\Upsilon}} &= \frac{1}{1-\sigma\mu(e^\star)^2}\left(\sigma\mu(e^\star)^2(\boldsymbol{\Upsilon}\cdot\mathbf{k})(\mathbf{k}\times\boldsymbol{\Upsilon}) + e^\star f_z(\mathbf{e}_z\times\mathbf{k})\right), \\ \dot{\mathbf{k}} &= \sigma\boldsymbol{\Upsilon}\times\mathbf{k}, \end{aligned} \tag{6.3}$$

where there is no tangential surface force. Clearly, this system is Hamiltonian, i.e., it conserves energy. It is readily shown that

$$\dot{\mathbf{v}_Q} = 0 \implies \dot{n} = 0 \implies \dot{\Upsilon}_z = 0 \implies \Upsilon\cdot(\mathbf{e}_z\times\mathbf{k}) = 0 \implies \dot{x} = \dot{y} = 0.$$

Solutions are therefore defined by level sets of $l^2 + m^2$ and $(\Upsilon_x)^2 + (\Upsilon_y)^2$ [6]. For $n \neq \pm 1$ these solutions are described by precession, spin about the axis of symmetry, and no nutation. These are Ebenfeld's tumbling solutions [6].

The abundance of invariant quantities in the no-slip, no-force problem severely limits the behavior of solutions to (6.3). In fact, the analysis above shows that the angular momentum about the vertical and the nutation angle fully determine solutions of (6.3). The nutation angle θ is the angle which the axis of symmetry makes with the vertical $\mathbf{e}_z$:

$$n = \cos(\theta).$$

We reasonably suppose that the initially sliding, viscous tippe top will tend to an asymptotic state, where Υ_z and n extremize the energy subject to the constraints in the problem. Thus, we suppose that $\aleph$ is the set of all energy-conserving asymptotic solutions defined by Υ_z and n which extremize the constrained energy.

6.1. Energy-momentum minimization. We now determine all minima of the total energy subject to the angular momentum constraint, $\Upsilon_Q = -\sigma(1+e^\star)$ (cf. (3.10)), and the attitude constraint, $\mathbf{k}\cdot\mathbf{k} = 1$. In other words, we find the extrema of the augmented energy $h = E + \lambda\Upsilon_Q + \hat{\lambda}\mathbf{k}\cdot\mathbf{k}$, where λ and $\hat{\lambda}$ are Lagrange multipliers. Extrema of h satisfy

$$\dot{x} = \dot{y} = \dot{z} = 0,$$

which indicates that the center of mass and nutation angle remain fixed. The angular momentum in terms of n takes the form

$$\boldsymbol{\Upsilon} = \frac{\lambda}{\sigma^2}\left(((\sigma-1)n + e^\star\sigma)\mathbf{k} + \mathbf{e}_z\right),$$

where

$$\lambda = \frac{(1+e^\star)\sigma^2}{(-1+n^2)(-1+\mu(e^\star)^2)+(n+e^\star)^2\sigma}.$$

Once the angular momentum takes the above form, the slip velocity vanishes: $\mathbf{v}_Q = 0$. This result is expected since extrema of the energy should have the same properties as the no-slip, no-force problem. With an angular momentum of this form, extrema of the energy satisfy

$$\begin{aligned}&\frac{\lambda^2}{\sigma^2}\left[(\mu(e^\star)^2-1)((\sigma-1)n+e^\star\sigma)^2\right]\mathbf{k}+2\hat{\lambda}\mathbf{k}\\&+\frac{\lambda^2}{\sigma^2}\left[(1-\sigma)n-e^\star\sigma-\mu(e^\star)^2n\right]\mathbf{e}_z+\mu e^\star\mathrm{Fr}^{-1}\mathbf{e}_z=0.\end{aligned}$$

When $\mathbf{e}_z$ and $\mathbf{k}$ are linearly dependent, we obtain the solutions $n = \pm 1$ and $\boldsymbol{\Upsilon} = \Upsilon_z\mathbf{e}_z$ with $\Upsilon_z = 1, \frac{1+e^\star}{1-e^\star}$, which correspond to the inverted and noninverted states, respectively. When the two vectors are linearly independent, we can choose $\hat{\lambda}$ so that the coefficient of $\mathbf{k}$ vanishes. The coefficient of $\mathbf{e}_z$ vanishes when the following polynomial equation in n is satisfied:

$$\left[(1-\sigma)n-e^\star\sigma-\mu(e^\star)^2n+\frac{\sigma^2}{\lambda^2}\mu e^\star\mathrm{Fr}^{-1}\right]=0.$$

This reduction to a one-dimensional problem in terms of components of the energy that are independent of the friction factor ν, and in terms of quantities in the plane defined by the vertical and the axis of symmetry, agrees with Ebenfeld's result [6]. This equation can be rewritten as a quartic polynomial of the form

$$a_0n^4+a_1n^3+a_2n^2+a_3n+a_4=0, \tag{6.4}$$

where the coefficients are functions of the parameters μ, Fr^{-1}, σ, and $e^\star$.

Heteroclinic connection existence criteria. We are not interested in explicitly solving the quartic equation (6.4). Rather, we are interested in knowing if all of the roots of this equation are outside the unit disc in the complex plane, since if $|n| > 1$, the extrema violates the constraint $\mathbf{k}\cdot\mathbf{k} = 1$. We can determine whether all roots satisfy this criterion by the following transformation in the complex domain:

$$n = \frac{-1+z}{1+z},$$

which describes a mapping from the region outside the open disc $|n| > 1$ to the open left half-plane in z. In terms of z the quartic polynomial becomes

$$r_0z^4+r_1z^3+r_2z^2+r_3z+r_4=0.$$

Specific values of these coefficients are provided in the appendix.

We now invoke the Liénard–Chipart criterion for the stability of quartic polynomials [7, p. 221].

Theorem 6.2 (Liénard–Chipart criterion). *Necessary and sufficient conditions for all the roots of $r_0 z^4 + r_1 z^3 + r_2 z^2 + r_3 z + a_4 = 0$ to have negative real parts can be given by*

$$r_0 > 0, \quad r_1 > 0, \quad r_2 > 0, \quad r_3 > 0, \quad r_4 > 0, \quad d = r_1 r_2 r_3 - r_0 r_3^2 - r_4 r_1^2 > 0.$$

In Figures 6.1, 6.2, and 6.3, we show the values of the coefficients and d for the cases $\mathrm{Fr}^{-1} = 0.3, 0.2, 0.1$, $\sigma = 1$, $\mu = 2.28$, and variable $e^\star$. Comparing these figures with the linear-stability plot for the noninverted and inverted equilibria in Figure 7.3, we observe that there are no other extrema in which the noninverted state is asymptotically unstable and the inverted state is asymptotically stable. In particular, for $\mathrm{Fr}^{-1} = 0.3, 0.2, 0.1$ and $e^\star < 0$, we observe that there are no other extrema in these ranges, respectively: $-0.1 < e^\star < 0.0$, $-0.2 < e^\star < 0.0$, and $-0.35 < e^\star < 0$. By evaluating the energy at the noninverted and inverted states, it can be shown that in these parameter ranges the inverted and noninverted states are absolute minima and maxima of the constrained energy, respectively.

In this regime of parameter values, the inverted state is the only point that remains in $\aleph$. By the theorems of Barbashin and Krasovskii, the inverted state is globally asymptotically stable in the parameter ranges, where the only extrema are the inverted and noninverted

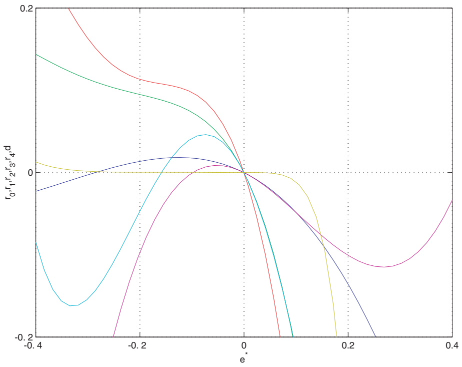


Figure 6.1. *Here we plot the coefficients of the quartic in $z = \frac{n+1}{1-n}$ for $\mu = 2.28$, $Fr^{-1} = 0.3$, $\sigma = 1.0$, and variable $e^\star$.*

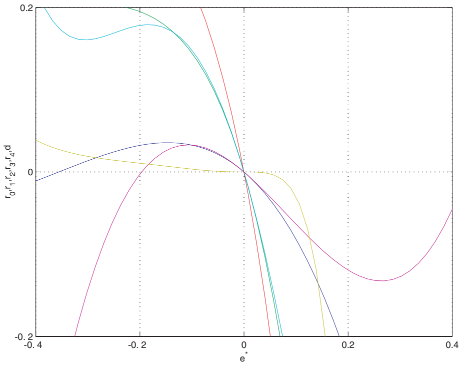


Figure 6.2. *Here we plot the coefficients of the quartic in $z = \frac{n+1}{1-n}$ for $\mu = 2.28$, $Fr^{-1} = 0.2$, $\sigma = 1.0$, and variable $e^\star$.*

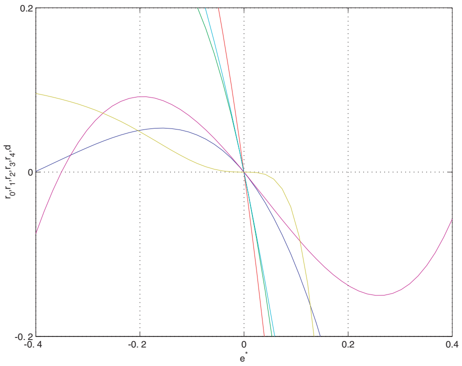


Figure 6.3. *Here we plot the coefficients of the quartic in* $z = \frac{n+1}{1-n}$ *for* $\mu = 2.28$*,* $Fr^{-1} = 0.1$*,* $\sigma = 1.0$*, and variable* $e^{\star}$*.*

states [8]. When the extrema include the tumbling solution, where $1 - n^2 \neq 0$, the heteroclinic connection does not exist. Ebenfeld provides specific criteria for the Lyapunov stability of this limit cycle. By the properties of the asymptotic states of the no-slip, no-force problem, it is clear this limit cycle should be neutrally stable.

The correspondence between the linear and nonlinear results is much more profound. We observe that the Hurwitz coefficients, which dictate stability of the polynomial in z, and hence determine the allowable nutation angles of extrema, are always r_0 and r_4. The appendix provides explicit expressions for these terms. Comparing the expressions for r_0 and r_4 with the stability criteria for the tippe top modified Maxwell–Bloch system equations (5.3) and (5.4), we observe that r_0 and r_4 are constant multiples of the third inequalities in (5.3) and (5.4). In section 5, we showed that these inequalities determine the linear stability of the noninverted and inverted states of the tippe top. Thus, *because of dissipation the linear analysis supplies all of the information needed to determine existence of the heteroclinic connection.*

7. Simulation.

We first reproduce some results in the literature by numerical integration of the governing equations (3.7). We time-integrate the equations of motion using the adaptive Runge–Kutta method. Plots of the nutation angle, precessional velocity, and spin velocity of the ball from Cohen's pioneering work are shown in Figure 7.1. The spin velocity figure illustrates conservation of angular momentum as the spin velocity changes direction when the axis of symmetry becomes horizontal. Figure 7.2 shows plots of nutation angle, spin velocity, and total energy as functions of time from Or's seminal work. The plot of the total, rotational, translational, and potential energies revealed that rotational, gravitational, and dissipative effects dominate translational effects.

We will now discuss data from simulations which confirm the local and global stability analysis in this paper. We consider tippe top inversion and tumbling for three cases: $\mathrm{Fr}^{-1} = 0.1$ (high Froude number), $\mathrm{Fr}^{-1} = 0.2$ (moderate Froude number), and $\mathrm{Fr}^{-1} = 0.3$ (low Froude number). Other parameters are valued at $\mu = 2.28$, $\sigma = 1.0$, and $\nu = 0.5$. Initial conditions for inversion are $\Upsilon_x = \Upsilon_y = 0.0$, $\Upsilon_z = 1.0$, $l = m = 0.0676$, $n = 0.9954$, $x = y = -0.0101$, $\dot{x} = 0.1010$, and $\dot{y} = 0.0309$. Initial conditions for tumbling are $\Upsilon_x = \Upsilon_y = 0.0$, $\Upsilon_z = (1.0 + e^{\star})/(1.0 - e^{\star}), 1.0$, $l = m = 0.0676$, $n = \mp 0.9954$, $x = y = -0.0101$, $\dot{x} = 0.1010$, and $\dot{y} = 0.0309$.

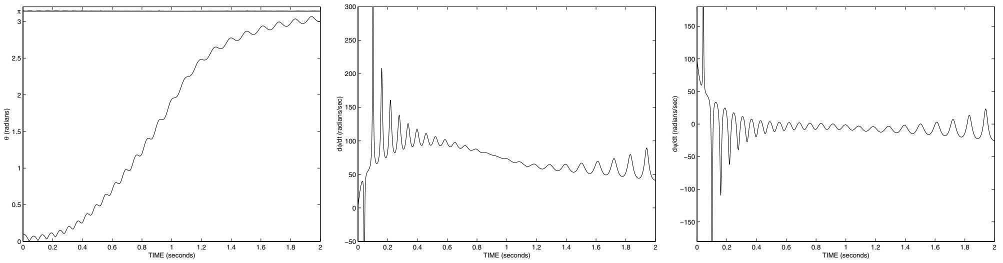


Figure 7.1. *From left: we show the nutation angle, the precessional velocity, and the spin rate as functions of time from Cohen's pioneering work* [5].

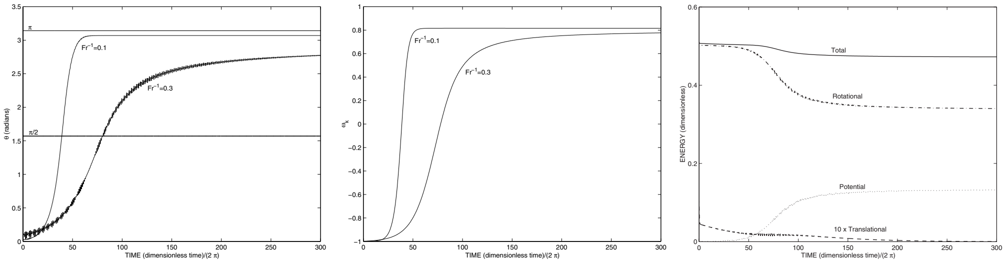


Figure 7.2. *From left: we show the nutation angle, the spin velocity, and the total energy and its components as functions of time from Or's seminal work* [11].

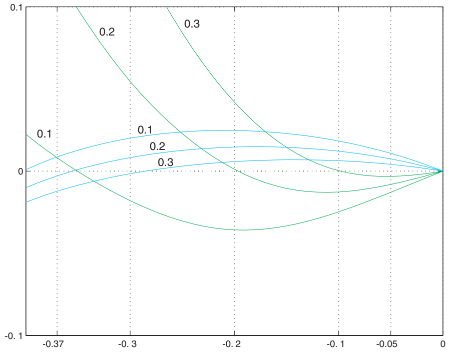


Figure 7.3. *Here we plot the growth factor of the eigenvalue of* (4.3) *with largest real part for both equilibria: the cyan curve corresponds to the noninverted state ($n^o = 1$, $\Upsilon_z{}^o = 1$), and the green curve corresponds to the inverted state ($n^o = -1$, $\Upsilon_z{}^o = \frac{1+e^\star}{1-e^\star}$) for varying Froude numbers ($Fr^{-1} = 0.1, 0.2, 0.3$). We use this stability plot to select appropriate parameter values for simulation of inversion and tumbling.*

The center-of-mass offset values $e^\star$ are chosen so that the noninverted state is asymptotically unstable and the inverted state is asymptotically stable (for inversion) and asymptotically unstable (for tumbling). We determine these values of $e^\star$ using the linear-stability plot in Figure 7.3. For low, moderate, and high Froude numbers, we pick $e^\star = (-0.05, -0.2)$, $(-0.1, -0.3), (-0.2, -0.37)$, respectively. The first $e^\star$ in the ordered pair corresponds to tippe top inversion and the second to tumbling. Figures 7.4, 7.6, and 7.8 show trajectories of the axis of symmetry on the unit ball for tippe top inversion. Figures 7.5, 7.7, and 7.9 show

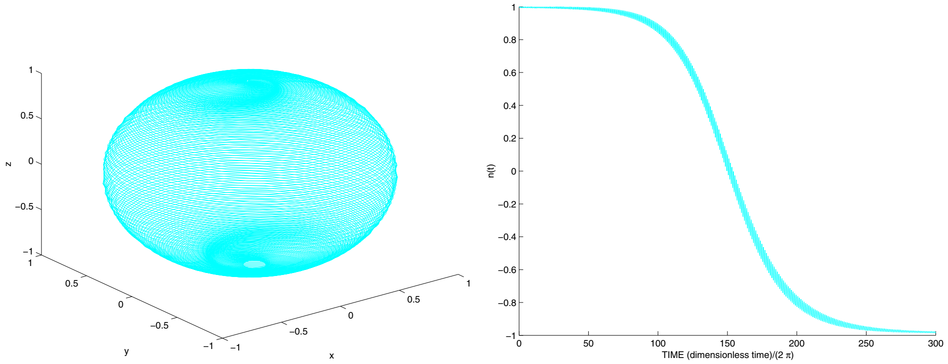


Figure 7.4. *Here we show the heteroclinic connection between the asymptotically unstable and stable states of the tippe top for* $Fr^{-1} = 0.1$*. Specifically, we show the trajectory of the axis of symmetry for a case when* χ*, the largest invariant set, is the isolated, asymptotically stable equilibrium point. Clicking on the above image displays the associated movie* (60135_01.mpg)*.*

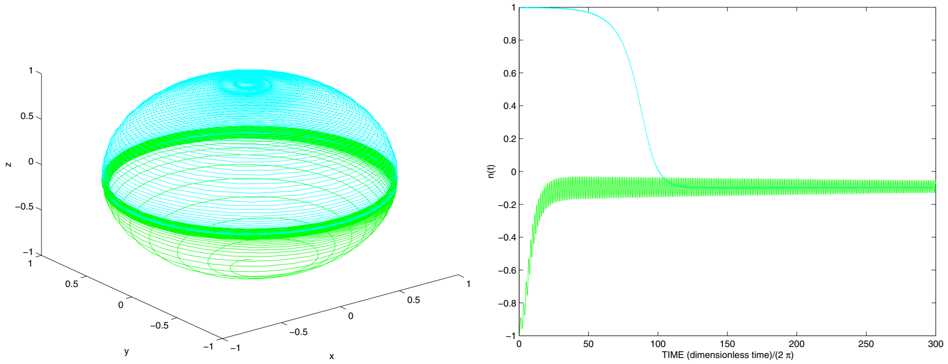


Figure 7.5. *This figure shows the trajectory of the axis of symmetry for a case when* χ *are only limit cycles defined by curves of constant energy, nutation angle, and angular momentum about the vertical for* $Fr^{-1} = 0.1$*. The parameter values for this case yield asymptotically unstable inverted and noninverted states. Clicking on the above image displays the associated movie* (60135_02.mpg)*.*

trajectories of the axis of symmetry on the unit ball for tippe top tumbling. Like Or, we observe that when the Froude number is small (weak spin relative to gravity) the tippe top takes longer to invert. As the Froude number increases, we observe that the tippe top flips more rapidly.

Six animations that correspond to each case above and that illustrate the phenomenon of tippe top inversion and tumbling can be found at http://www.acm.caltech.edu/~nawaf/tippetop.html. The animations show the evolution of the axis of symmetry and center of mass. It is very clear from these animations that the tippe top rotates much more than it translates: numerical evidence that translation of the mass center is not necessary to understand the linearized behavior of the ball.

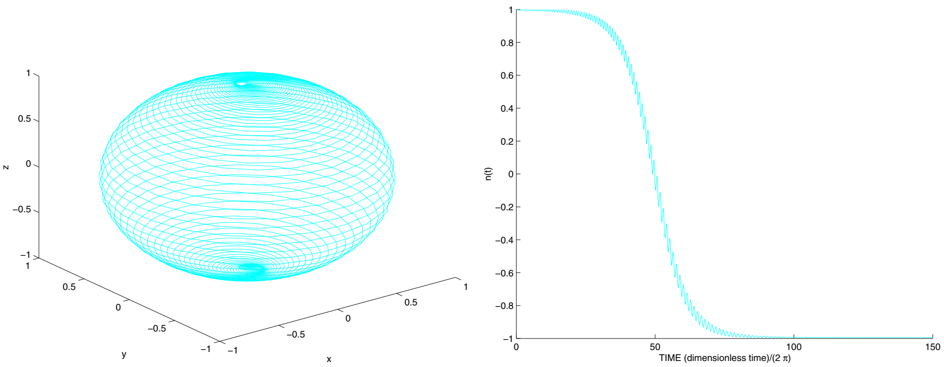


Figure 7.6. *Here we show the heteroclinic connection between the asymptotically unstable and stable states of the tippe top for $Fr^{-1} = 0.2$. Specifically, we show the trajectory of the axis of symmetry for a case when χ, the largest invariant set, is the isolated, asymptotically stable equilibrium point. Clicking on the above image displays the associated movie* (60135_03.mpg).

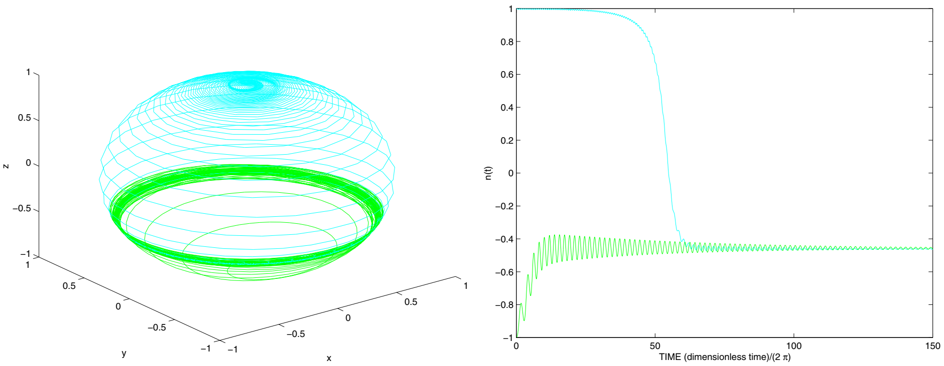


Figure 7.7. *This figure shows the trajectory of the axis of symmetry for a case when χ are only limit cycles defined by curves of constant energy, nutation angle, and angular momentum about the vertical for $Fr^{-1} = 0.2$. The parameter values for this case yield asymptotically unstable inverted and noninverted states. Clicking on the above image displays the associated movie* (60135_04.mpg).

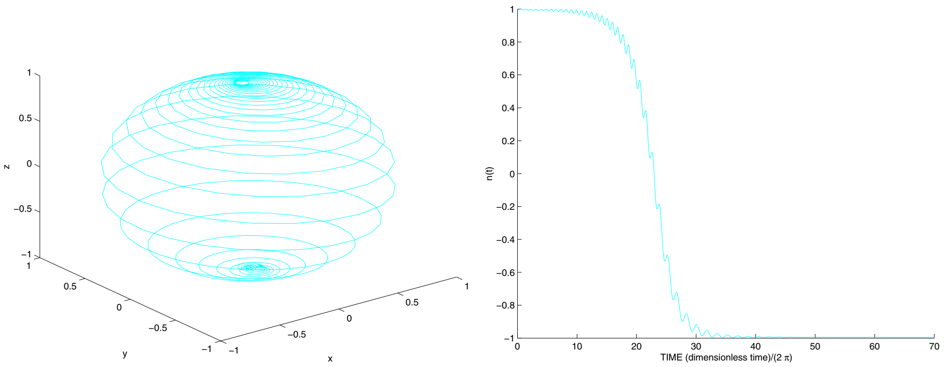


Figure 7.8. *Here we show the heteroclinic connection between the asymptotically unstable and stable states of the tippe top for $Fr^{-1} = 0.3$. Specifically, we show the trajectory of the axis of symmetry for a case when χ, the largest invariant set, is the isolated, asymptotically stable equilibrium point. Clicking on the above image displays the associated movie* (60135_05.mpg).

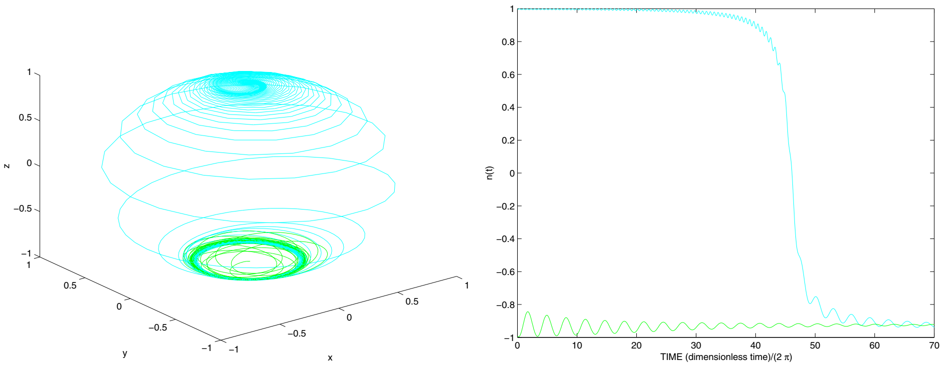


Figure 7.9. *This figure shows the trajectory of the axis of symmetry for a case when* χ *are only limit cycles defined by curves of constant energy, nutation angle, and angular momentum about the vertical for* $Fr^{-1} = 0.3$*. The parameter values for this case yield asymptotically unstable inverted and noninverted states. Clicking on the above image displays the associated movie* (60135_06.mpg).

8. Future directions.

Possible future directions include investigating other axisymmetric physical systems described by the modified Maxwell–Bloch equations (e.g., rotating beam and levitron). For a discussion of the related, but different, rising egg, please see Bou-Rabee, Marsden, and Romero [3].

Appendix.

Normal reaction force.

An explicit expression for the surface normal reaction force follows:

$$\begin{aligned} f_z =& \frac{\mu e^\star \sigma}{1-\mu(e^\star)^2 n^2}\left((\boldsymbol{\Upsilon}\cdot\mathbf{k})\Upsilon_z - (\boldsymbol{\Upsilon}\cdot\boldsymbol{\Upsilon})(1-\mu(e^\star)^2)n - \mu(e^\star)^2 n(\boldsymbol{\Upsilon}\cdot\mathbf{k})^2\right) \\ &+ \frac{\mu e^\star}{1-\mu(e^\star)^2 n^2}(1+e^\star n)(\mathbf{k}\cdot\mathbf{f}_f) + \frac{\mu(1-\mu(e^\star)^2)}{\sigma(1-\mu(e^\star)^2 n^2)}\mathrm{Fr}^{-1}. \end{aligned}$$

Governing equations in component form.

The attitude equations are

$$\begin{aligned} \dot{l} &= (-\Upsilon_z m + \Upsilon_y n)\sigma, \\ \dot{m} &= (\Upsilon_z l - \Upsilon_x n)\sigma, \\ \dot{n} &= (-\Upsilon_y l + \Upsilon_x m)\sigma. \end{aligned}$$

The translational equations are

$$\begin{aligned} \ddot{x} &= -\frac{\nu}{\mu}(\dot{x} - e^\star \dot{l} + (\sigma-1)(\boldsymbol{\Upsilon}\cdot\mathbf{k}) - \sigma\Upsilon_y) = \frac{\sigma}{\mu} f_x, \\ \ddot{y} &= -\frac{\nu}{\mu}(\dot{y} - e^\star \dot{m} - (\sigma-1)(\boldsymbol{\Upsilon}\cdot\mathbf{k}) + \sigma\Upsilon_x) = \frac{\sigma}{\mu} f_y. \end{aligned}$$

The rotational equations are

$$
\begin{aligned}
\dot{\Upsilon}_x &= \frac{1}{\sigma(1-\mu(e^\star)^2)}(-\mu(e^\star)^2\sigma(\boldsymbol{\Upsilon}\cdot\mathbf{k})\dot{l} - e^\star\sigma f_z m + \sigma(e^\star n+1)f_y - \mu(e^\star)^2\sigma(-mf_x+lf_y)l),\\
\dot{\Upsilon}_y &= \frac{1}{\sigma(1-\mu(e^\star)^2)}(-\mu(e^\star)^2\sigma(\boldsymbol{\Upsilon}\cdot\mathbf{k})\dot{m} + e^\star\sigma f_z l - \sigma(e^\star n+1)f_x - \mu(e^\star)^2\sigma(-mf_x+lf_y)m),\\
\dot{\Upsilon}_z &= \frac{1}{\sigma(1-\mu(e^\star)^2)}(-\mu(e^\star)^2\sigma(\boldsymbol{\Upsilon}\cdot\mathbf{k})\dot{n} - e^\star\sigma(-mf_x+lf_y)(1+\mu e^\star n)).
\end{aligned}
$$

Hurwitz coefficients. The values of the parameters in the polynomial

$$r_0z^4 + r_1z^3 + r_2z^2 + r_3z + a_4 = 0,$$

where $z = \frac{n+1}{1-n}$, are

$$
\begin{aligned}
r_0 &= (1+e^\star)^2\sigma^2(1+\mathrm{Fr}^{-1}e^\star(1+e^\star)^2\mu - \sigma - e^\star(e^\star\mu+\sigma)),\\
r_1 &= -2(1+e^\star)^2\sigma(2\mathrm{Fr}^{-1}e^\star\mu(-2+(e^\star)^2(2\mu-\sigma)+\sigma) + \sigma(-1+(e^\star)^2\mu+\sigma+2e^\star\sigma)),\\
r_2 &= 2e^\star(-3(1+e^\star)^2\sigma^3 + \mathrm{Fr}^{-1}\mu(8(-1+(e^\star)^2\mu)^2\\
&\quad - 8(-1+(e^\star)^2)(-1+(e^\star)^2\mu)\sigma + 3(-1+(e^\star)^2)^2\sigma^2)),\\
r_3 &= -2\sigma(2\mathrm{Fr}^{-1}(-1+e^\star)^2e^\star\mu(-2+(e^\star)^2(2\mu-\sigma)+\sigma) - (1+e^\star)^2\sigma(-1+(e^\star)^2\mu+\sigma-2e^\star\sigma)),\\
r_4 &= \sigma^2(-1+\sigma+e^\star(-2+\mathrm{Fr}^{-1}\mu+\sigma+e^\star(-1+((1+e^\star)^2\\
&\quad + \mathrm{Fr}^{-1}(-2+e^\star)(2+(-2+e^\star)e^\star))\mu - \sigma - e^\star\sigma))).
\end{aligned}
$$

Acknowledgment. We wish to acknowledge Andy Ruina for helpful comments.

REFERENCES

[1] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and T. S. Ratiu, *Dissipation induced instabilities*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 37–90.

[2] A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and T. S. Ratiu, *The Euler–Poincaré equations and double bracket dissipation*, Comm. Math. Phys., 175 (1996), pp. 1–42.

[3] N. M. Bou-Rabee, J. E. Marsden, and L. A. Romero, *A geometric treatment of Jellett's egg*, ZAMM, submitted, 2004.

[4] M. G. Clerc and J. E. Marsden, *Dissipation-induced instabilities in an optical cavity laser: A mechanical analog near the* 1:1 *resonance*, Phys. Rev. E, 64 (2001), 067603.

[5] R. J. Cohen, *The tippe top revisited*, Amer. J. Phys., 45(1) (1977), pp. 12–17.

[6] S. Ebenfeld and F. Scheck, *A new analysis of the tippe top: Asymptotic states and Liapunov stability*, Ann. Phys., 243 (1995), pp. 195–217.

[7] F. R. Gantmacher, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1959.

[8] H. K. Khalil, *Nonlinear Systems*, 3rd ed., Prentice–Hall, Upper Saddle River, NJ, 2002, pp. 126–133.

[9] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.

[10] H. K. Moffatt and Y. Shimomura, *Classical dynamics: Spinning eggs—a paradox resolved*, Nature, 416 (2002), pp. 385–386.

[11] A. C. Or, *The dynamics of a tippe top*, SIAM J. Appl. Math., 54 (1994), pp. 597–609.

[12] E. J. Routh, *The Advanced Part of a Treatise on the Dynamics of a System of Rigid Bodies*, Macmillan, New York, 1905.

[13] A. Ruina, *Rolling and Sliding of Spinning Things: Euler's Disk, Jellett's Egg and Moffatt's Nature.* http://tam.cornell.edu/~ruina/hplab/Rolling%20and%20sliding/index.html (2002).

[14] W. Thomson and P. G. Tait, *Treatise on Natural Philosophy*, Cambridge University Press, Cambridge, UK, 1879.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 3, pp. 378–407

Breathing Pulses in an Excitatory Neural Network[*]

Stefanos E. Folias[†] and Paul C. Bressloff[†]

Abstract. In this paper we show how a local inhomogeneous input can stabilize a stationary-pulse solution in an excitatory neural network. A subsequent reduction of the input amplitude can then induce a Hopf instability of the stationary solution resulting in the formation of a breather. The breather can itself undergo a secondary instability leading to the periodic emission of traveling waves. In one dimension such waves consist of pairs of counterpropagating pulses, whereas in two dimensions the waves are circular target patterns.

Key words. traveling waves, excitatory neural network, breathers, integro-differential equations

AMS subject classification. 92C20

DOI. 10.1137/030602629

1. Introduction.

Various in vitro experimental studies have observed waves of excitation propagating in cortical slices when stimulated appropriately [7, 12, 35]. The propagation velocity of these synaptically generated waves is of the order 0.06m/s, which is much slower than the typical speed of 0.5m/s found for action-potential propagation along axons. Traveling waves of electrical activity have also been observed in vivo in the somatosensory cortex of behaving rats [24], turtle and mollusk olfactory bulbs [18, 21], turtle cortex [27], and visuomotor cortices in the cat [29]. Often these traveling waves occur during periods without sensory stimulation, with the subsequent presentation of a stimulus inducing a switch to synchronous oscillatory behavior [11]. This suggests that determining the conditions under which cortical wave propagation can occur is important for understanding the normal processing of sensory stimuli as well as more pathological forms of behavior such as epileptic seizures and migraines.

A number of theoretical studies have established the occurrence of traveling fronts [8, 15] and traveling pulses [36, 1, 19, 25, 30] in one-dimensional excitatory neural networks modeled in terms of nonlinear integro-differential equations. Such equations are infinite dimensional dynamical systems and can be written in the general form [10]

$$
\begin{aligned}
\frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} w(x|x')f(u(x',t))dx' - \beta q(x,t) + I(x),\\
\frac{1}{\epsilon}\frac{\partial q(x,t)}{\partial t} &= -q(x,t) + u(x,t),
\end{aligned}
\tag{1.1}
$$

where $u(x,t)$ is a neural field that represents the local activity of a population of excitatory neurons at position $x \in \mathbf{R}$, $I(x)$ is an external input current, $f(u)$ denotes the output firing

[*] Received by the editors December 19, 2003; accepted for publication (in revised form) by D. Terman April 14, 2004; published electronically September 29, 2004.
http://www.siam.org/journals/siads/3-3/60262.html

[†] Department of Mathematics, University of Utah, 155 S 1400 E, Salt Lake City, UT 84112 (sfolias@math.utah.edu, bressloff@math.utah.edu).

rate function, and $w(x|x')$ is the strength of connections from neurons at x' to neurons at x. The neural field $q(x,t)$ represents some form of negative feedback mechanism such as spike frequency adaptation or synaptic depression, with β, ε determining the relative strength and rate of feedback. The nonlinear function f is typically taken to be a sigmoid function $f(u) = 1/(1 + \mathrm{e}^{-\gamma(u-\kappa)})$ with gain γ and threshold κ. It can be shown [25] that there is a direct link between the above model and experimental studies of wave propagation in cortical slices where synaptic inhibition is pharmacologically blocked [7, 12, 35]. Since there is strong vertical coupling between cortical layers, it is possible to treat a thin cortical slice as an effective one-dimensional medium. Analysis of the model provides valuable information regarding how the speed of a traveling wave, which is relatively straightforward to measure experimentally, depends on various features of the underlying cortical circuitry.

One of the common assumptions in the analysis of traveling wave solutions of (1.1) is that the system is spatially homogeneous; that is, the external input $I(x)$ is independent of x and the synaptic weights depend only on the distance between presynaptic and postsynaptic cells, $w(x|x') = w(x - x')$. The existence of traveling waves can then be established for a class of weight distributions $w(x)$ that includes the exponential function $\mathrm{e}^{-|x|/d}$. The waves are in the form of traveling fronts in the absence of any feedback [8], whereas traveling pulses tend to occur when there is significant feedback [25]. The real cortex, however, is more realistically modeled as an inhomogeneous medium. Inhomogeneities in the synaptic weight distribution w may arise due to the patchy nature of long-range horizontal connections in superficial layers of cortex. For example, in the primary visual cortex the horizontal connections tend to link cells with similar stimulus feature preferences such as orientation and ocular dominance [23, 37, 3]. The variation of the feature preferences across the cortex is approximately periodic, and this induces a corresponding periodic modulation in the horizontal connections. It has previously been shown that an inhomogeneous periodic modulation in the strength of synaptic interactions induced by long-range patchy connections can lead to wavefront propagation failure [4]. If the wavelength of the periodic inhomogeneity is much shorter than the characteristic wavelength of the front, then averaging theory can be used to achieve an effective homogenization of the neural medium along similar lines to that previously developed for a model of calcium waves [16] and for a model of chemical waves in a bistable medium [17].

Another important source of spatial inhomogeneity is the external input $I(x)$. Such inputs would arise naturally from sensory stimuli in the case of the intact cortex and could be introduced by external stimulation in the case of cortical slices. We have recently shown how a monotonically varying input can induce wave propagation failure due to the pinning of a stationary-front solution [5, 6]. More significantly, the stationary front can subsequently destabilize via a Hopf bifurcation as the degree of input inhomogeneity is reduced, resulting in an oscillatory back-and-forth pattern of wave propagation. Analogous *breather*-like front solutions have previously been found in inhomogeneous reaction-diffusion systems [28, 31, 13, 14, 2, 26] and in numerical simulations of a realistic model of fertilization calcium waves [22].

In this paper we extend our work on fronts by analyzing the effects of input inhomogeneities on the stability of stationary pulses, since these better reflect the types of neural activity patterns observed in the cortex. In order to construct exact wave solutions, we follow previous treatments [1, 25] and consider the high gain limit $\gamma \to \infty$ of the sigmoid function f such that $f(u) = H(u - \kappa)$, where H is the Heaviside step function; that is, $H(u) = 1$ if $u \geq 0$ and

$H(u) = 0$ if $u < 0$. As a further simplification, we also assume that the weight distribution w is homogeneous so that (1.1) reduces to the form

$$\begin{aligned}\frac{\partial u(x,t)}{\partial t} &= -u(x,t) + \int_{-\infty}^{\infty} w(x-x')H(u(x',t)-\kappa)dx' - \beta q(x,t) + I(x),\\ \frac{1}{\epsilon}\frac{\partial q(x,t)}{\partial t} &= -q(x,t) + u(x,t).\end{aligned} \tag{1.2}$$

We first construct explicit traveling wave solutions of (1.2) in the case of a constant input (section 2). We then analyze the existence and stability of stationary pulses in the presence of a unimodal input (section 3). We show that (i) a sufficiently large input inhomogeneity can stabilize a stationary pulse and (ii) a subsequent reduction in the level of inhomogeneity can induce a Hopf instability of the stationary pulse leading to the formation of a breather-like oscillatory wave. Numerically we find that a secondary instability can occur beyond which the breather periodically emits pairs of traveling pulses (section 4). Moreover, there is mode-locking between the oscillation frequency of the breather and the rate of wave emission. Analogous forms of oscillatory wave are also shown to occur in a more biophysically realistic conductance-based model (section 4). Finally, we extend our analysis to radially symmetric pulses in a two-dimensional network (section 5).

2. Traveling pulses in a homogeneous network.

We begin by briefly outlining the construction of traveling pulse solutions of (1.2) in the case of zero input $I(x) = 0$, following the approach of Pinto and Ermentrout [25]. We assume that the weight distribution $w(x)$ is a positive even function of x, is a monotonically decreasing function of $|x|$, and satisfies the normalization condition $\int_{-\infty}^{\infty} w(x)dx < \infty$. For concreteness, w is taken to be an exponential weight distribution

$$w(x) = \frac{1}{2d}\mathrm{e}^{-|x|/d}. \tag{2.1}$$

The length scale is fixed by setting $d = 1$. Translation symmetry implies that we can consider traveling pulse solutions of the form $u(x,t) = U(x - ct)$ with $U(\pm\infty) = 0$ and $U(-a) = U(0) = \kappa$. Without loss of generality, we take $c > 0$. Substituting into (1.2) with $I(x) = 0$ and differentiating with respect to ξ lead to the second-order boundary value problem

$$\begin{gathered}-c^2U''(\xi) + c[1+\epsilon]U'(\xi) - \epsilon[1+\beta]U(\xi) = c[w(\xi+a) - w(\xi)] + \epsilon[W(\xi+a) - W(\xi)],\\ U(0) = U(-a) = \kappa,\\ U(\pm\infty) = 0,\end{gathered} \tag{2.2}$$

where $W(\xi) = \int_0^{\xi} w(x)dx$. Equation (2.2) is solved by considering separately the domains $\xi \le -a$, $-a \le \xi \le 0$, and $\xi \ge 0$ and matching the solutions at $\xi = -a, 0$. On the domain $\xi > 0$, with w given by the exponential function (2.1),

$$\begin{gathered}-c^2U''(\xi) + c(1+\epsilon)U'(\xi) - \epsilon(1+\beta)U(\xi) = \frac{c+\epsilon}{2}\left(\mathrm{e}^{-\xi} - \mathrm{e}^{-(\xi+a)}\right),\\ U(0) = \kappa,\\ U(\infty) = 0.\end{gathered} \tag{2.3}$$

This has the solution $U_>(\xi) = \kappa \mathrm{e}^{-\xi}$ provided that

$$\kappa = \frac{(c+\epsilon)(1-\mathrm{e}^{-a})}{2(c^2 + c(1+\epsilon) + \epsilon(1+\beta))} \equiv f(c,a). \tag{2.4}$$

On the other two domains we have solutions consisting of complementary and particular parts:

$$U_0(\xi) = \mathcal{A}_+ \mathrm{e}^{\mu_+ \xi} + \mathcal{A}_- \mathrm{e}^{\mu_- \xi} + \mathcal{U}_+ \mathrm{e}^{\xi} + \mathcal{U}_- \mathrm{e}^{-\xi} \tag{2.5}$$

for $-a < \xi < 0$ and

$$U_<(\xi) = \mathcal{A}'_+ \mathrm{e}^{\mu_+ \xi} + \mathcal{A}'_- \mathrm{e}^{\mu_- \xi} + \mathcal{U}'_+ \mathrm{e}^{\xi} \tag{2.6}$$

for $\xi < -a$, where

$$\mu_\pm = \frac{1}{2c}\left[1 + \epsilon \pm \sqrt{(1+\epsilon)^2 - 4\epsilon(1+\beta)}\right]. \tag{2.7}$$

The coefficients $\mathcal{U}_\pm$ and $\mathcal{U}'_+$ are obtained by direct substitution into the differential equation for U, whereas the four coefficients $\mathcal{A}_\pm$ and $\mathcal{A}'_\pm$ are determined by matching solutions at the boundaries. This leads to the five boundary conditions (i) $U_0(0) = \kappa$, (ii) $U_0(-a) = \kappa$, (iii) $U_<(-a) = \kappa$, (iv) $U'_0(0) = -\kappa$, and (v) $U'_0(-a) = U'_<(-a)$. Since there are five equations in four unknowns, we generate a second constraint on the speed and size of the wave that is of the form $g(c,a) = \kappa$. A full solution to the wave equation can then be found for just those values of c, a which satisfy both $f(c,a) = \kappa$ and $g(c,a) = \kappa$.

Pinto and Ermentrout [25] used a shooting method to show that for sufficiently slow negative feedback (small ϵ) and large β there exist two pulse solutions, one narrow and slow and the other wide and fast. Numerically, the fast solution is found to be stable [25]. It is also possible to construct an explicit stationary-pulse solution by setting $c = 0$ in (2.2):

$$U(\xi) = \begin{cases} \dfrac{\mathrm{e}^{-\xi}}{2(1+\beta)}\left(1 - \mathrm{e}^{-a}\right) & \text{for } \xi > 0, \\ \dfrac{1}{2(1+\beta)}\left[2 - \mathrm{e}^{\xi} - \mathrm{e}^{-(\xi+a)}\right] & \text{for } -a < \xi < 0, \\ \dfrac{\mathrm{e}^{\xi}}{2(1+\beta)}\left(\mathrm{e}^{a} - 1\right) & \text{for } \xi < -a \end{cases} \tag{2.8}$$

with

$$\kappa = \frac{1}{2(1+\beta)}\left(1 - \mathrm{e}^{-a}\right). \tag{2.9}$$

It turns out that stationary-pulse solutions are unstable in the case of homogeneous inputs (see section 3), acting as separatrices between a zero activity state and a traveling wavefront.

3. Stationary pulses in an inhomogeneous network.

We now investigate the existence and stability of one-dimensional stationary pulses in the presence of a unimodal input $I(x)$ which, for concreteness, is taken to be a Gaussian of width σ centered at the origin

$$I(x) = \mathcal{I} \mathrm{e}^{-x^2/2\sigma^2}. \tag{3.1}$$

As in section 2, we take w to be a positive even function of x and a monotonically decreasing function of $|x|$, and we choose the normalization $\int_{-\infty}^{\infty} w(x)dx = 1$. For illustrative purposes, the exponential weight distribution (2.1) will be used as a specific example.

3.1. Stationary-pulse existence.

From symmetry arguments there exists a stationary-pulse solution $U(x)$ of (1.2) centered at $x = 0$, satisfying

$$\begin{array}{lll} U(x) > \kappa, & x \in (-a, a); & U(\pm a) = \kappa, \\ U(x) < \kappa, & x \in (-\infty, -a) \cup (a, \infty); & U(\pm\infty) = 0. \end{array}$$

In particular,

$$(1 + \beta)\, U(x) = \int_{-a}^{a} w(x - x')dx' + I(x). \tag{3.2}$$

The threshold κ and width a are related according to the self-consistency condition

$$\hat{\kappa} = [I(a) + W(2a)] \equiv G(a), \tag{3.3}$$

where $\hat{\kappa} = (1+\beta)\kappa$ and $W(2a) = \int_0^{2a} w(x)dx$. The existence or otherwise of a stationary-pulse solution can then be established by finding solutions to (3.3). Consider, for example, the exponential weight distribution (2.1) with $d = 1$ such that $W(2a) = (1 - \mathrm{e}^{-2a})/2$. Furthermore, suppose that the amplitude $\mathcal{I}$ of the Gaussian input (3.1) is treated as a bifurcation parameter with the range σ kept fixed. (The effect of varying σ will be discussed below.) It is straightforward to show that there always exists a critical amplitude $\mathcal{I}_c$, below which $G(a)$ is strictly monotonically increasing and above which $G(a)$ has two stationary points. Consequently, as $\hat{\kappa}$ varies, we have the possibility of zero, one, two, or three stationary-pulse solutions. The function $G(a)$ is plotted in Figure 1 for a range of input amplitudes $\mathcal{I}$, with horizontal lines indicating different values of $\hat{\kappa}$: intersection points determine the existence of stationary-pulse solutions. Let κ_c denote the value of $G(a)$ for which $G'(a)$ has a double zero. Anticipating the stability results of section 3.2, we obtain the following results. If $\hat{\kappa} < \kappa_c$, then there is only a single pulse solution branch which is always unstable. On the other hand, if $\hat{\kappa} > \kappa_c$, then there are two distinct bifurcation scenarios (see Figure 2), both of which can support a stable pulse solution.

Scenario (i): $\kappa_c < \hat{\kappa} < 1/2$. There exist three solution branches with the lower (narrow pulse) and upper (wide pulse) branches unstable. If $\epsilon > \beta$, then the middle (intermediate pulse) branch is stable along its entire length, annihilating in a saddle-node bifurcation at the endpoints S, S'. On the other hand, if $\epsilon < \beta$, then only a central portion of the middle branch is stable due to the existence of two Hopf bifurcation points H, H'. In the limit $\epsilon \to \beta$ we have $H \to S$ and $H' \to S'$ leading to some form of degenerate bifurcation. Note that as $\widehat{\kappa} \to 1/2$, $a_{S'} \to \infty$, thus causing the upper branch to collapse.

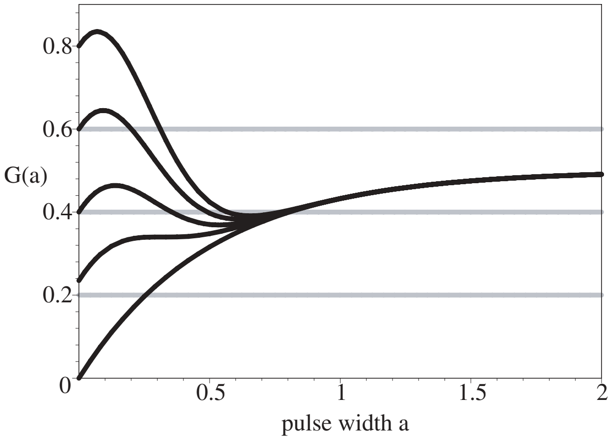


Figure 1. *Plot of $G(a)$ in (3.3) as a function of pulse width a for an exponential weight distribution and various values of input amplitude $\mathcal{I}$ with $\sigma = 0.25$. Horizontal lines (gray) represent different values of $\hat{\kappa} = \kappa(1+\beta)$. Intersections of black and gray curves indicate the existence of stationary-pulse solutions.*

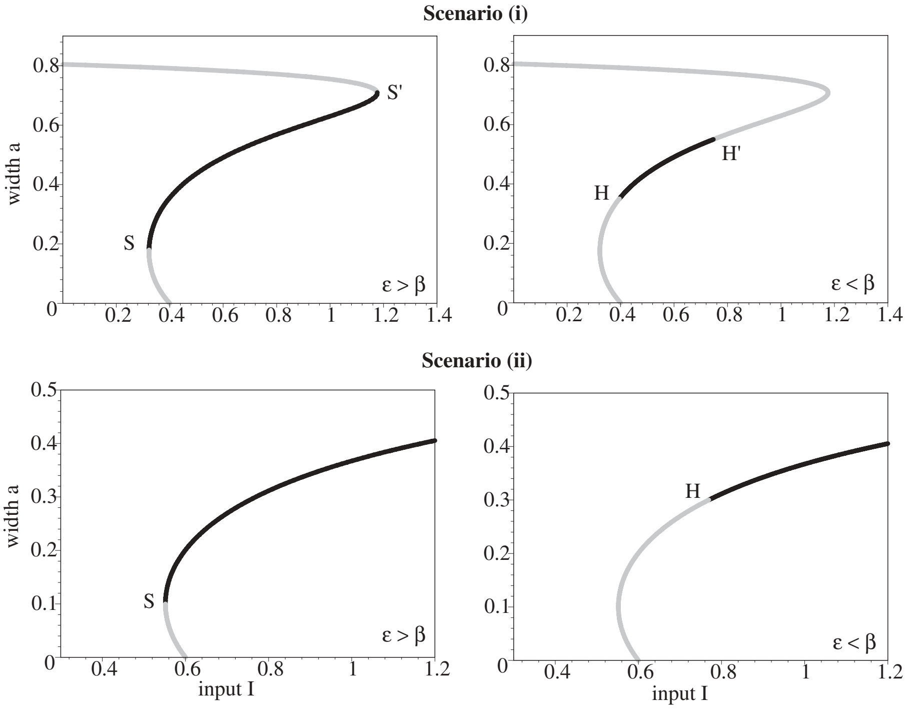


Figure 2. *One-dimensional stationary-pulse existence curves for an exponential weight distribution and* (i) $\kappa_c < \hat{\kappa} < \frac{1}{2}$, (ii) $\hat{\kappa} > \frac{1}{2}$. *Other parameter values are $\beta = 1$, $\sigma = 0.25$. Black indicates stability, whereas gray indicates instability of the stationary pulse. Saddle-node bifurcation points are indicated by S, S' and Hopf bifurcation points by H, H'.*

Scenario (ii): $\hat{\kappa} > 1/2$. There exist two solution branches with the lower branch unstable and the upper branch stable for sufficiently large $\mathcal{I}$. If $\epsilon > \beta$, then the upper branch is stable along its entire length, annihilating in a saddle-node bifurcation at its endpoint S. On the other hand, if $\epsilon < \beta$, then the upper branch loses stability via a Hopf bifurcation at the point H with $H \to S$ as $\epsilon \to \beta$.

In both of the above scenarios there also exists a stable subthreshold solution $U(x) = I(x)/(1+\beta)$ when $\mathcal{I} < \hat{\kappa}$. This is coexistent with the lower suprathreshold pulse, and the pair annihilate at $\mathcal{I} = \hat{\kappa}$. To address the effect of varying the input σ, consider the case where $\hat{\kappa} < \frac{1}{2}$. As σ decreases, κ_c decreases, widening the $\hat{\kappa}$-interval for which there exist three stationary-pulse solutions: in particular, $\kappa_c \to 0$ as $\sigma \to 0$. Conversely, as σ increases, κ_c increases toward $\frac{1}{2}$, thus decreasing the size of the three-pulse regime. For $\hat{\kappa} > \frac{1}{2}$, qualitatively, the bifurcation scenario remains unchanged; the effect of increasing σ is simply to widen the pulse width a. Finally, note that the qualitative behavior of the function $G(a)$, which determines the existence of stationary-pulse solutions, follows from the fact that both $w(x)$ and $I(x)$ are monotonically decreasing functions of $|x|$ and are symmetric about $x = 0$.

3.2. Stability analysis. The stability of a stationary pulse of width a is determined by writing $u(x,t) = U(x) + \varphi(x,t)$ and $q(x,t) = Q(x) + \psi(x,t)$ with $Q(x) = U(x)$ and expanding (1.2) to first-order in (φ, ψ). This leads to the linear equation

$$\begin{aligned} \frac{\partial \varphi(x,t)}{\partial t} &= -\varphi(x,t) + \int_{-\infty}^{\infty} w(x-x')H'(U(x'))\varphi(x',t)dx' - \beta\psi(x,t), \\ \frac{1}{\epsilon}\frac{\partial \psi(x,t)}{\partial t} &= -\psi(x,t) + \varphi(x,t). \end{aligned} \tag{3.4}$$

We assume that $\varphi, \psi \in L^1(\mathbf{R})$. The spectrum of the associated linear operator is found by taking $\varphi(x,t) = \mathrm{e}^{\lambda t}\varphi(x)$ and $\psi(x,t) = \mathrm{e}^{\lambda t}\psi(x)$ and using the identity

$$\frac{dH(U(x))}{dU} = \frac{\delta(x-a)}{|U'(a)|} + \frac{\delta(x+a)}{|U'(-a)|}, \tag{3.5}$$

where

$$U'(x) = \frac{1}{1+\beta}\left[I'(a) + w(x+a) - w(x-a)\right] \tag{3.6}$$

and $U'(-a) = -U'(a) > 0$. We then obtain the eigenvalue equation

$$\left(\lambda + 1 + \frac{\epsilon\beta}{\lambda+\epsilon}\right)\varphi(x) = \frac{w(x+a)}{|U'(-a)|}\varphi(-a) + \frac{w(x-a)}{|U'(a)|}\varphi(a). \tag{3.7}$$

Note that we have formally differentiated the Heaviside function, which is permissible since it arises inside a convolution. One could also develop the linear stability analysis by considering perturbations of the threshold crossing points along the lines of Amari [1]. Since we are linearizing about a stationary rather than a traveling pulse, we can analyze the spectrum of the linear operator without the recourse to Evans functions.

There exist three types of solutions to (3.7). The first consists of functions $\varphi(x)$ that vanish at $x = \pm a$ and $\lambda = \lambda_\pm^{(0)}$ with $\lambda_\pm^{(0)}$ given by

$$\lambda_\pm^{(0)} = \frac{-(1+\epsilon) \pm \sqrt{(1+\epsilon)^2 - 4\epsilon(1+\beta)}}{2}. \tag{3.8}$$

Note that $\lambda_\pm^{(0)}$ belong to the essential spectrum since they have infinite multiplicity and that they are always real and negative. Thus they do not contribute to instabilities. The second consists of solutions of the form $\varphi(x) = A[w(x+a) - w(x-a)]$ with λ given by the roots of the equation

$$\lambda + 1 + \frac{\epsilon\beta}{\lambda+\epsilon} = \frac{w(0) - w(2a)}{|U'(a)|}. \tag{3.9}$$

It follows that $\lambda = \lambda_\pm$, where

$$\lambda_\pm = \frac{-\Lambda \pm \sqrt{\Lambda^2 - 4(1-\Gamma)\epsilon(1+\beta)}}{2} \tag{3.10}$$

with

$$\Lambda = 1 + \epsilon - (1+\beta)\Gamma, \quad \Gamma = \frac{w(0) - w(2a)}{w(0) - w(2a) + D}, \tag{3.11}$$

and $D = |I'(a)|$. Finally, the third type of solution is $\varphi(x) = A[w(x+a) + w(x-a)]$ with λ given by the roots of the equation

$$\lambda + 1 + \frac{\epsilon\beta}{\lambda+\epsilon} = \frac{w(0) + w(2a)}{|U'(a)|}. \tag{3.12}$$

This yields $\lambda = \widehat{\lambda}_\pm$, where

$$\widehat{\lambda}_\pm = \frac{-\widehat{\Lambda} \pm \sqrt{\widehat{\Lambda}^2 - 4(1-\widehat{\Gamma})\epsilon(1+\beta)}}{2} \tag{3.13}$$

with

$$\widehat{\Lambda} = 1 + \epsilon - (1+\beta)\widehat{\Gamma}, \quad \widehat{\Gamma} = \frac{w(0) + w(2a)}{w(0) - w(2a) + D}. \tag{3.14}$$

A stationary-pulse solution will be stable provided that $\operatorname{Re}\lambda_\pm, \operatorname{Re}\widehat{\lambda}_\pm < 0$.

In the limiting case of a homogeneous input, for which $D = 0$, (3.10) and (3.13) become

$$\lambda_- = 0, \quad \lambda_+ = \beta - \epsilon, \tag{3.15}$$

and

$$\widehat{\lambda}_\pm = \frac{-\widehat{\Lambda}_0 \pm \sqrt{\widehat{\Lambda}_0^2 + 4\epsilon(1+\beta)(\widehat{\Gamma}_0 - 1)}}{2} \tag{3.16}$$

with

$$\widehat{\Lambda}_0 = \epsilon + 1 - (1+\beta)\widehat{\Gamma}_0, \tag{3.17}$$

$$\widehat{\Gamma}_0 = \frac{w(0)+w(2a)}{w(0)-w(2a)}. \tag{3.18}$$

Since $\widehat{\Gamma}_0 > 1$ for finite pulse width a, it follows that $\widehat{\lambda}_+ > 0$ for all parameter values, and, hence, a stationary pulse (if it exists) is unstable in the case of the homogeneous network described by (1.2). This result is consistent with Amari's previous analysis [1]. He showed that in order to stabilize a stationary pulse within a homogeneous network, it is necessary to include some form of lateral inhibition. If a weak input inhomogeneity is subsequently introduced into the network, then the peak of the activity profile moves to a local maximum of the input where it is pinned.

For a nonzero Gaussian input, the gradient of a stationary pulse is given by $D(a)$, where

$$D(a) = \frac{a\mathcal{I}}{\sigma^2}\mathrm{e}^{-(a/\sigma)^2/2}. \tag{3.19}$$

Using the gradient, we wish to determine the stability of the pulse in terms of the pulse width a, with $a = a(\mathcal{I})$ given by one of the solutions of (3.3) for fixed κ, β. Stability of the stationary pulse corresponds to the following conditions:

$$\Gamma, \hat{\Gamma} < 1, \qquad \Lambda, \hat{\Lambda} > 0.$$

However, there are redundancies. First, by inspection of (3.11), the condition $\Gamma < 1$ is automatically satisfied. The conditions $\Lambda, \hat{\Lambda} > 0$ are equivalent to

$$\Gamma, \hat{\Gamma} < \frac{1+\epsilon}{1+\beta},$$

and, since $\Gamma < \hat{\Gamma}$, it follows that the condition on Γ is redundant. Hence, stability of the stationary pulse reduces to the conditions

$$\hat{\Gamma} < 1, \qquad \hat{\Gamma} < \frac{1+\epsilon}{1+\beta},$$

in which the latter is redundant for $\epsilon > \beta$, while the former is redundant for $\epsilon < \beta$. These conditions translate in terms of the gradient D as

$$\epsilon > \beta: \qquad D(a) > 2w(2a) \equiv D_{\mathbf{SN}}(a), \tag{3.20}$$

$$\epsilon < \beta: \qquad D(a) > D_c(a), \tag{3.21}$$

where

$$D_c(a) = 2w(2a) + \left(\frac{\beta-\epsilon}{1+\epsilon}\right)(w(0)+w(2a)). \tag{3.22}$$

We now relate stability of the stationary pulse to the gradient D on different branches of the existence curves shown in Figure 2 for $w(x)$ given by the exponential distribution (2.1).

Stability for $\epsilon > \beta$. Equation (3.3) implies that $D(a) = 2w(2a) - G'(a)$. Thus, stability condition (3.20) is satisfied when $G'(a) < 0$ and not satisfied when $G'(a) > 0$. Saddle-node bifurcation points occur when $G'(a) = 0$, i.e., when $D(a)$ passes through D_{SN}, due to the vanishing of a single real eigenvalue $\hat{\lambda}_+$. We can make the following conclusions about the solution branches. In scenario (i) there are three solution branches. On the lower and upper solution branches, $G'(a) > 0$, while $G'(a) < 0$ on the middle branch, indicating that the former are always unstable and that the latter is stable for $\epsilon > \beta$. In scenario (ii) there are two solution branches: using the same arguments, the lower branch is always unstable while, for $\epsilon > \beta$, the upper branch is stable.

Hopf curves for $\epsilon < \beta$. If $\epsilon < \beta$, then a Hopf bifurcation can occur due to a complex pair of eigenvalues $\hat{\lambda}_\pm$ crossing into the right half complex plane. The Hopf bifurcation point is determined by the condition $\hat{\Gamma} = (1+\varepsilon)/(1+\beta) < 1$, which is equivalent to the gradient condition $D(a) = D_c(a) > D_{\text{SN}}(a)$. It follows that only branches determined to be stable for $\epsilon > \beta$ can undergo a Hopf bifurcation when $\epsilon < \beta$. Moreover, the Hopf bifurcation points coincide with saddle-node bifurcation points precisely at the point $\beta = \epsilon$, where there is a pair of zero eigenvalues suggestive of a codimension 2 Takens–Bogdanov bifurcation. As ϵ decreases from β, we expect the Hopf bifurcation point(s) to traverse these previously stable branches from the saddle-node point(s). In order to illustrate this, we find a relationship for $D(a)$ which does not depend explicitly on $\mathcal{I}$. Using (3.3), the input gradient D can be related as

$$\begin{aligned} D(a) &= |I'(a)| \\ &= \frac{a}{\sigma^2} I(a) \\ &= \frac{a}{\sigma^2}\left(\kappa(1+\beta) - W(2a)\right). \end{aligned} \tag{3.23}$$

We restrict a here depending on which branch of the existence curve we are considering. In each of the scenarios discussed in section 3.1, we examine graphically the crossings of the curves $D(a), D_c(a)$: stability corresponds to $D(a) > D_c(a)$ with Hopf points at $D(a) = D_c(a)$. Figure 3 illustrates the generic behavior in these scenarios. The left column presents the graphs of D and of D_c for different values of ϵ spanning the interval $[0, \beta]$; intersection points indicate Hopf bifurcation points. The right column graphs the corresponding Hopf curves in (a, ϵ)-parameter space. Note that the upper branch in scenario (ii) is always stable for sufficiently large input $\mathcal{I}$, that is, for large pulse width a, for (3.3) implies that $I(a) \sim (1+\beta)\kappa - 1/2$, and hence $D \sim [(1+\beta)\kappa - 1/2](a/\sigma^2)$ as $a \to \infty$. Since $\widehat{D}_c(a) \to (\beta - \epsilon)/(1+\epsilon)$ and $\mathrm{e}^{-2a} \to 0$ as $a \to \infty$, it follows that both stability conditions (3.20) and (3.21) are satisfied in this limit. Varying σ does not affect the qualitative behavior of the Hopf bifurcation curves. Since σ appears only in (3.23), the effect of increasing σ is to shrink the graph of D by a factor $1/\sigma^2$, causing the Hopf curves in the right column of Figure 3 to be stretched downward, thus increasing the size of the stability region in the (a, ϵ)-plane.

4. Numerical results.

In our numerical simulations we use a Runge–Kutta (RK4) scheme with 4000–10000 spatial grid points and time step $dt = 0.02$, evaluating the integral term by

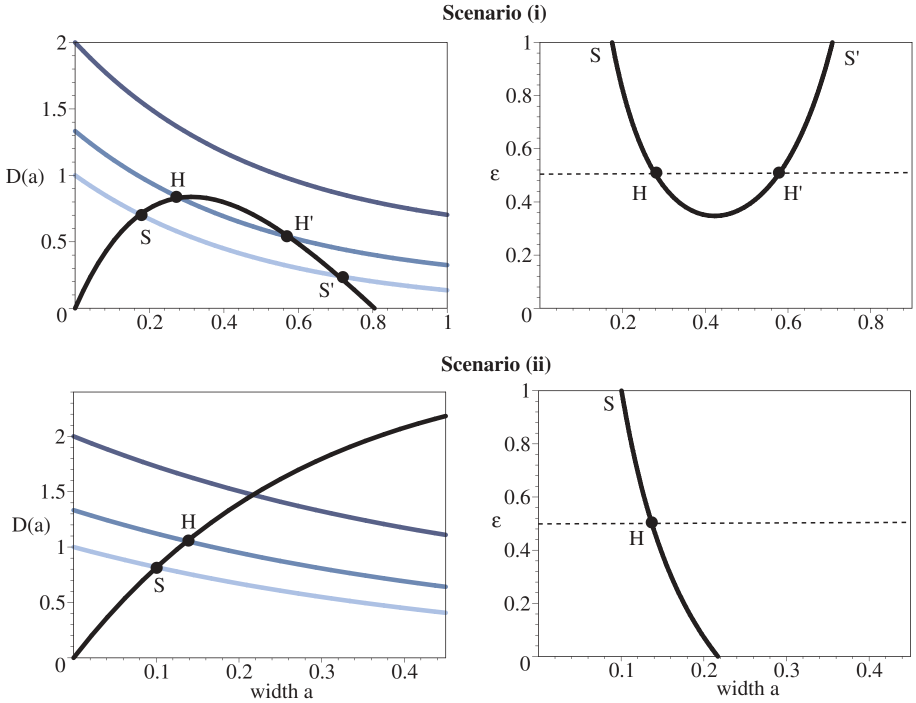


Figure 3. *Left column: Gradient curves for the two bifurcation scenarios shown in Figure* 2: (i) $\kappa_c < \hat{\kappa} < \frac{1}{2}$ *and* (ii) $\hat{\kappa} > \frac{1}{2}$. *The thick solid curve shows the input gradient $D(a)$ as a function of pulse width a. The lighter curves show the critical gradient $D_c(a)$ as function of a for $\epsilon = 0.0, 0.5, 1.0$ and $\beta = 1$. For a given value of $\epsilon < \beta$, a stationary pulse of width a is stable provided that $D(a) > D_c(a)$. A pulse loses stability via a Hopf bifurcation at any intersection points $D(a) = D_c(a)$. The Hopf bifurcation point(s) for $\epsilon = 0.5$ are indicated by H, H'. In the limit $\epsilon \to \beta$, we have $H, H' \to S, S'$. Right column: Corresponding Hopf stability curves in the (a, ϵ)-plane.*

quadrature. Boundary points freely evolve according to the scheme rather than by prescription, and the size of the domain is chosen so that the stationary pulse is unaffected by the boundaries.

4.1. Hopf bifurcation to a breather. Numerically solving the one-dimensional rate equation (1.2), we find that the Hopf instability of the upper solution branch in bifurcation scenario (ii) induces a breather-like oscillatory pulse solution; see Figures 4 and 5. As the input amplitude $\mathcal{I}$ is slowly reduced below $\mathcal{I}_{HB}$, the oscillations steadily grow until a new instability point is reached. Interestingly, the breather persists over a range of inputs beyond this secondary instability, except that it now periodically emits pairs of traveling pulses, as illustrated in Figure 6. In fact, such a solution is capable of persisting even when the input is below threshold, that is, for $\mathcal{I} < (1+\beta)\kappa$. Note that although the homogeneous network ($\mathcal{I} = 0$) also supports the propagation of traveling pulses, it does not support the existence of a breather that can act as a source of these waves.

Our simulations suggest both supercritical and subcritical Hopf bifurcations can occur for scenario (ii). The conclusion of supercriticality is based on the evidence that there is continuous growth of the amplitude of the oscillations from the stationary solution as $\mathcal{I}$ is

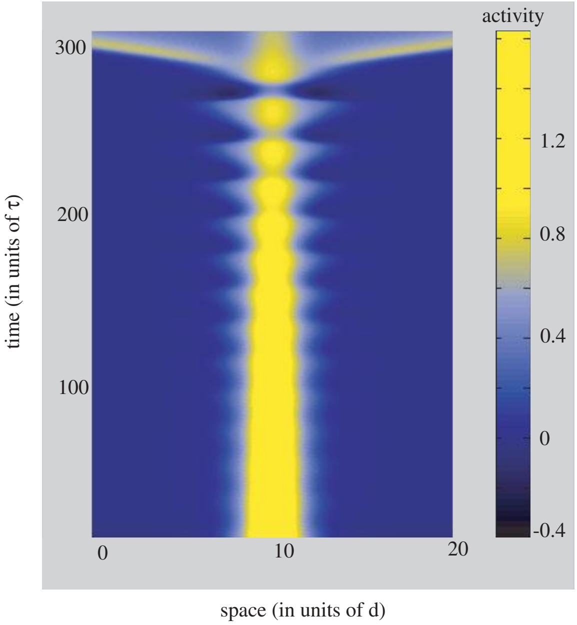


Figure 4. *Breather-like solution arising from a Hopf instability of a stationary pulse due to a slow reduction in the amplitude $\mathcal{I}$ of the Gaussian input inhomogeneity* (3.1) *for an exponential weight distribution. Here $\mathcal{I} = 5.5$ at $t = 0$ and $\mathcal{I} = 1.5$ at $t = 250$. Other parameter values are $\epsilon = 0.03$, $\beta = 2.5$, $\kappa = 0.3$, $\sigma = 1.0$. The amplitude of the oscillation steadily grows until it undergoes a secondary instability at $\mathcal{I} \approx 2$, beyond which the breather persists and periodically generates pairs of traveling pulses (only one of which is shown). The breather itself disappears when $\mathcal{I} \approx 1$.*

reduced through the predicted bifurcation point, and, moreover, that the frequency of the oscillatory solution near the bifurcation point is approximately equal to the predicted Hopf frequency

$$\omega_H = \operatorname{Im} \hat{\lambda}_{\pm} = \sqrt{\epsilon(\beta - \epsilon)}.$$

For example, the Hopf bifurcation of the stationary pulse for the parameter values given in Figure 4 was determined numerically to be supercritical. Conversely, the Hopf bifurcation in scenario (i) appears to be subcritical. Furthermore, the basin of attraction of the stable pulse on the middle branch seems to be small, rendering it, as well as any potential breather, difficult to approach. Hence, we did not investigate this case further.

As mentioned above, a secondary instability occurs at some $\mathcal{I} < \mathcal{I}_{HB}$, whereupon traveling pulses are emitted: this behavior appears to occur only for values of ϵ that support traveling pulses in the homogeneous model ($\mathcal{I} = 0$). As the point of secondary instability is approached, the breather starts to exhibit behavior suggestive of pulse emission, except that the recovery variable q increases rapidly enough to prevent the nascent waves from propagating. On the other hand, beyond the point of instability, recovery is not fast enough to block pulse emission; we also find that the activity variable u always drops well below threshold after each emission. Interestingly, for a range of input amplitudes we observe frequency-locking between the oscillations of the breather and the rate at which pairs of pulses are emitted from the

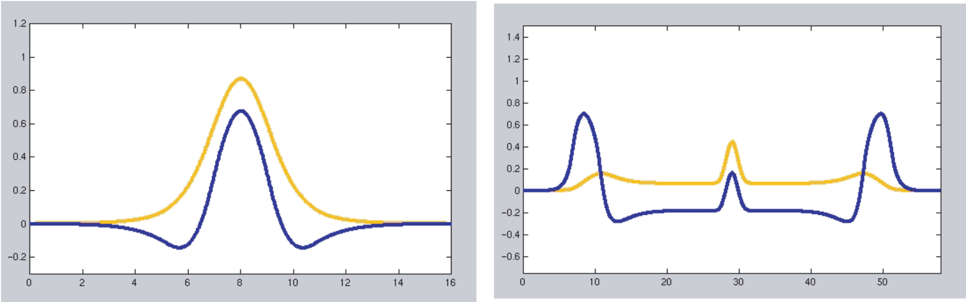


Figure 5. *Snapshots of the oscillatory behavior of* (a) *the breathing pulse* ($\mathcal{I} = 2.2$) *and* (b) *the pulse emitter* ($\mathcal{I} = 1.3$) *far beyond the bifurcation point. The graphs of U and Q are indicated in blue and gold, respectively, with the horizontal axis representing space in units of d, the spatial extent associated with the exponential weight function. Other parameters are $\beta = 2.5$, $\kappa = 0.3$, $\epsilon = 0.03$. Clicking on the above images displays the associated movies.*

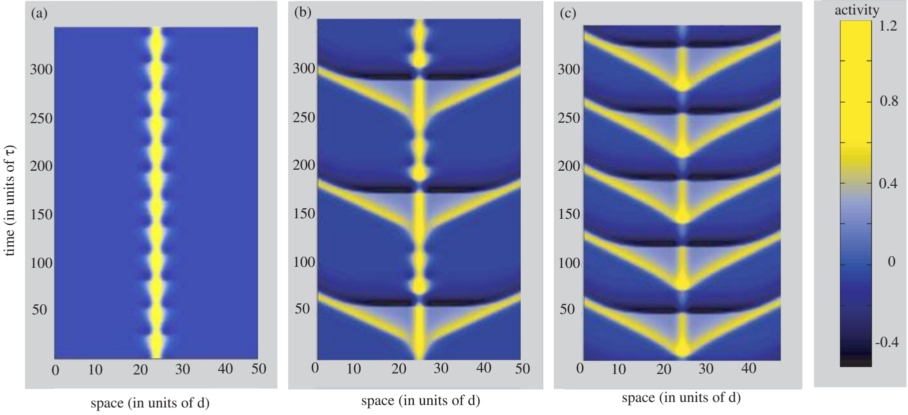


Figure 6. *Mode-locking in the transition from breather to pulse emitter.* (a) 0:1 *mode-locking for* $\mathcal{I} = 2.3$, (b) 1:4 *mode-locking for* $\mathcal{I} = 2.1$, (c) 1:2 *mode-locking for* $\mathcal{I} = 1.3$.

breather. Two examples of $n : m$ mode-locked solutions are shown in Figure 6, in which there are n pairs of pulses emitted per m oscillation cycles of the central breather. As $\mathcal{I}$ is reduced further, the only mode that is seen is 1:2, which itself ultimately vanishes, and the system is attracted to the subthreshold solution.

4.2. Breathers in a biophysical model. Although the rate model is very useful as an analytically tractable model of neural tissue, it is important to determine whether or not its predictions regarding spatio-temporal dynamics hold in more biophysically realistic conductance-based models. For concreteness, we consider a version of the Traub model, in which there is an additional slow potassium M-current that produces the effect of spike-rate adaptation [9]. We also discretize space by setting $x = j\Delta x$ for $j = 1, \dots, N$ and label neurons by the index j. The membrane potential of the jth neuron satisfies the following Hodgkin–Huxley-like dynamics [9]:

$$C\frac{dV_j}{dt} = -I_{\text{ion}}(V_j, m, n, h, q) - I_j^{\text{syn}}(t) + I_j$$

with synaptic current

$$\begin{aligned} I_j^{\text{syn}}(t) &= g_{\text{syn}} \sum_k w(|j-k|) s_k(t)(V - V_{\text{syn}}), \\ \frac{ds_j}{dt} &= K(V_j(t))(1 - s_j) - \frac{1}{\tau} s_j \end{aligned}$$

and ionic currents

$$I_{\text{ion}}(V, m, n, h) = g_{\text{L}}(V - V_{\text{L}}) + g_{\text{K}} n^4 (V - V_{\text{K}}) + g_{\text{Na}} m^3 h (V - V_{\text{Na}}) + g_q q (V - V_{\text{K}}),$$

$$\tau_p(V)\frac{dp}{dt} = p_\infty(V) - p, \qquad p \in \{m, n, h, q\}.$$

The various biophysical model functions and the parameters used in the numerics are listed in Appendix A. Note that we have also included an external Gaussian input current I_j in order to investigate the behavior predicted by the rate model. Without this external input, the biophysical model has previously been shown to support a traveling pulse, consisting of either a single action-potential or a packet of action-potentials [9]. Since the firing rate model describes the average activity, we interpret *high activity* as repetitive firing of neurons and *low activity* as neurons that are subthreshold or quiescent. Hence, we expect that the application of a strong unimodal input should generate a stationary pulse, i.e., a localized region of neurons that are repetitively firing, surrounded by a region of neurons that are quiescent. Subsequent reduction of the input should lead to oscillations in this localized region followed by emission of action-potentials or packets of action-potentials.

One obvious difference between this biophysical model and the rate model discussed in section 3 is that the gating variable associated with spike-rate adaptation evolves according to more complicated nonlinear dynamics, while that of the firing rate model evolves according to simple linear dynamics. Nevertheless, the behavior of the rate model appears to carry over to the biophysical model, thus lending support to the ability of rate models to describe the averaged behavior of spiking biophysical models. For large input amplitude $\mathcal{I}$, the system approaches a solution in which a region, localized about the input, is repetitively firing, while the outer region is quiescent; moreover, the firing rate is maximal in the center of this region and decreases toward the boundaries, which is analogous to the stationary pulse of the firing rate model. As $\mathcal{I}$ is subsequently decreased, there is a transition to breather-like behavior: periodically, packets or bursts of action-potentials begin to propagate from the active region and, shortly thereafter, fail to propagate as the newly excited region recovers. As in the rate model, further reduction of the amplitude $\mathcal{I}$ leads to a transition to a state in which packets of persistent action-potentials are emitted. Two examples are shown in Figure 7. The first is in a regime where the breather still dominates with the occasional emission of wave packets. The second corresponds to regular pulse emission, in which periodic bursts of persistent action-potentials are emitted, each followed by an interlude of subthreshold behavior in the vicinity

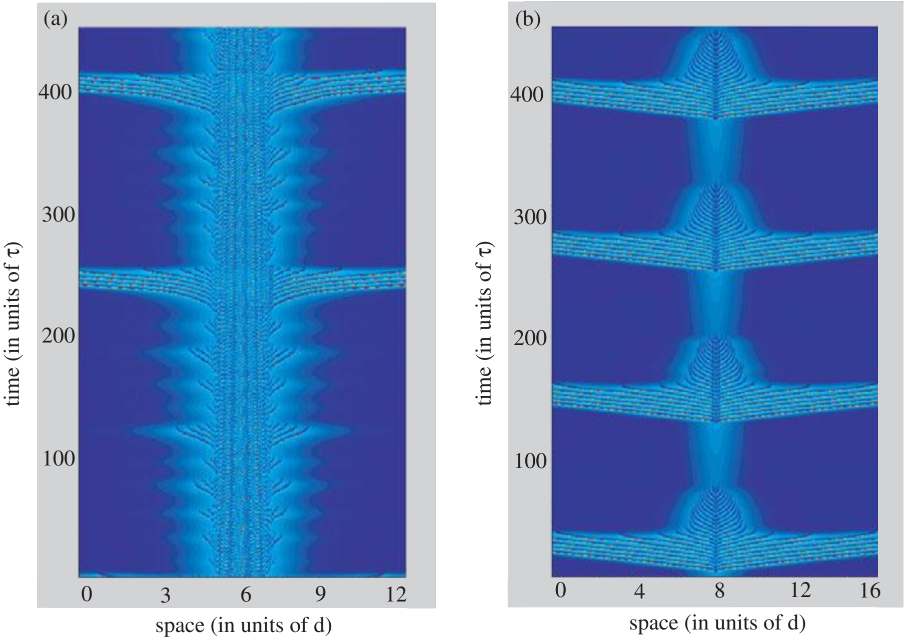


Figure 7. *Breathers in a biophysical model with an exponential weight distribution.* (a) $\mathcal{I} = 75\text{mA/cm}^2$, (b) $\mathcal{I} = 50\text{mA/cm}^2$, *where $\mathcal{I}$ is the amplitude of the Gaussian input. Other parameter values are specified in Appendix* A.

of the input; this is similar to the 1:2 pulse emitter of the firing rate model shown in Figure 6. We expect similar behavior to occur in other biophysical models that have some form of sufficiently slow negative feedback.

5. Two-dimensional pulses.

We now extend our analysis to derive conditions for the existence and stability of radially symmetric stationary-pulse solutions of a two-dimensional version of (1.2):

$$\begin{aligned} \frac{\partial u(\mathbf{r},t)}{\partial t} &= -u(\mathbf{r},t) + \int_{\mathbb{R}^2} w(|\mathbf{r}-\mathbf{r}'|)H(u(\mathbf{r}',t)-\kappa)d\mathbf{r}' - \beta q(\mathbf{r},t) + I(r), \\ \frac{1}{\epsilon}\frac{\partial q(\mathbf{r},t)}{\partial t} &= -q(\mathbf{r},t) + u(\mathbf{r},t), \end{aligned} \tag{5.1}$$

where $\mathbf{r} = (r,\theta)$ and $\mathbf{r}' = (r',\theta')$. Both the input $I(r)$ and the weight distribution $w(r)$ are taken to be positive monotonically decreasing functions in $L^1(\mathbf{R}^+)$. As in the one-dimensional case, stationary-pulse solutions are unstable in a homogeneous excitatory network but can be stabilized by the local input. Our analysis should be contrasted with a number of recent studies of two-dimensional stationary pulses [32, 34, 20]. These latter studies consider homogeneous networks with uniform external inputs and include both excitatory and inhibitory synaptic coupling. Thus the inhibition is nonlocal rather than local as in our model.

5.1. Stationary-pulse existence.

We begin by developing a formal representation of the two-dimensional stationary-pulse solution for a general monotonically decreasing weight function w. We then generate stationary-pulse existence curves for the specific case of an exponential weight function and analyze their dependence on the parameters of the system. Since we cannot obtain a closed form for the solution in the case of the exponential weight distribution, we also derive an explicit solution for the case of a modified Bessel weight function that approximates the exponential. For concreteness, we consider a Gaussian input $I(r) = \mathcal{I}\mathrm{e}^{-r^2/2\sigma^2}$.

Pulse construction for a general synaptic weight function. A radially symmetric stationary-pulse solution of (5.1) is $u = q = U(r)$ with U depending only upon the spatial variable r such that

$$\begin{aligned} &U(r) > \kappa, \quad r \in (0,a); && U(\infty) = 0,\\ &U(a) = \kappa; && U(0) < \infty,\\ &U(r) < \kappa, \quad r \in (a,\infty). && \end{aligned}$$

Substituting into (5.1) gives

$$(1+\beta)U(r) = M(a,r) + I(r), \tag{5.2}$$

where

$$\begin{aligned} M(a,r) &= \int_{\mathbf{R}^2} w(|\mathbf{r}-\mathbf{r}'|)H(U(r')-\kappa)d\mathbf{r}'\\ &= \int_0^{2\pi}\int_0^a w(|\mathbf{r}-\mathbf{r}'|)r'dr'd\theta. \end{aligned} \tag{5.3}$$

In order to calculate the double integral in (5.3) we use the Fourier transform, which for radially symmetric functions reduces to a Hankel transform. To see this, consider the two-dimensional Fourier transform of the radially symmetric weight function w, expressed in polar coordinates,

$$\begin{aligned} w(r) &= \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{i(\mathbf{r}\cdot\mathbf{k})}\breve{w}(\mathbf{k})d\mathbf{k}\\ &= \frac{1}{2\pi}\int_0^\infty \left(\int_0^{2\pi} e^{ir\rho\cos(\theta-\phi)}\breve{w}(\rho)d\phi\right)\rho d\rho, \end{aligned}$$

where $\breve{w}$ denotes the Fourier transform of w and $\mathbf{k} = (\rho,\phi)$. Using the integral representation

$$\frac{1}{2\pi}\int_0^{2\pi} e^{ir\rho\cos(\theta-\varphi)}d\theta = J_0(r\rho),$$

where $J_\nu(z)$ is the Bessel function of the first kind, we express w in terms of its Hankel transform of order zero,

$$w(r) = \int_0^\infty \breve{w}(\rho)J_0(r\rho)\rho d\rho, \tag{5.4}$$

which, when substituted into (5.3), gives

$$M(a,r) = \int_0^{2\pi}\int_0^a \left(\int_0^\infty \breve{w}(\rho)J_0(\rho|\mathbf{r}-\mathbf{r}'|)\rho d\rho\right) r'dr'd\theta'.$$

Switching the order of integration gives

$$M(a,r) = \int_0^\infty \breve{w}(\rho)\left(\int_0^{2\pi}\int_0^a J_0(\rho|\mathbf{r}-\mathbf{r}'|)r'dr'd\theta'\right)\rho d\rho. \tag{5.5}$$

In polar coordinates $|\mathbf{r}-\mathbf{r}'| = \sqrt{r^2+r'^2-2rr'\cos(\theta-\theta')}$,

$$\begin{aligned}\int_0^{2\pi}\int_0^a J_0(\rho|\mathbf{r}-\mathbf{r}'|)r'dr'd\theta' &= \int_0^{2\pi}\int_0^a J_0\left(\rho\sqrt{r^2+r'^2-2rr'\cos(\theta-\theta')}\right)r'dr'd\theta' \\ &= \frac{1}{\rho^2}\int_0^{2\pi}\int_0^{a\rho} J_0\left(\sqrt{R^2+R'^2-2RR'\cos(\theta')}\right)R'dR'd\theta',\end{aligned}$$

where $R = r\rho$ and $R' = r'\rho$. To separate variables, we use the addition theorem

$$J_0\left(\sqrt{R^2+R'^2-2RR'\cos\theta'}\right) = \sum_{m=0}^\infty \epsilon_m J_m(R)J_m(R')\cos m\theta',$$

where $\epsilon_0 = 1$ and $\epsilon_n = 2$ for $n \geq 1$. Since $\int_0^{2\pi}\cos m\theta' d\theta' = 0$ for $m \geq 1$, it follows that

$$\begin{aligned}\int_0^{2\pi}\int_0^a J_0(\rho|\mathbf{r}-\mathbf{r}'|)r'dr'd\theta' &= \frac{1}{\rho^2}\sum_{m=0}^\infty \epsilon_m J_m(R)\int_0^{a\rho} J_m(R')R'dR'\int_0^{2\pi}\cos m\theta'd\theta' \\ &= \frac{2\pi}{\rho^2}J_0(R)\int_0^{a\rho} J_0(R')R'dR' \\ &= \frac{2\pi a}{\rho}J_0(r\rho)J_1(a\rho).\end{aligned}$$

Hence for general weight w, $M(a,r)$ has the formal representation

$$M(a,r) = 2\pi a\int_0^\infty \breve{w}(\rho)J_0(r\rho)J_1(a\rho)d\rho. \tag{5.6}$$

We now wish to show that for a general monotonically decreasing weight function $w(r)$, the function $M(a,r)$ is necessarily a monotonically decreasing function of r. This will ensure that the radially symmetric stationary-pulse solution (5.2) is also a monotonically decreasing function of r in the case of a Gaussian input. Differentiating M with respect to r using (5.3) yields

$$\frac{\partial M}{\partial r}(a,r) = \int_0^{2\pi}\int_0^a w'(|\mathbf{r}-\mathbf{r}'|)\left(\frac{r-r'\cos(\theta')}{\sqrt{r^2+r'^2-2rr'\cos(\theta')}}\right)r'dr'd\theta'. \tag{5.7}$$

By inspection of (5.7), $\frac{\partial M}{\partial r}(a,r) < 0$ for $r > a$, since $w'(z) < 0$. To see that it is also negative for $r < a$ and, thus, monotonic, we instead consider the equivalent Hankel representation of (5.6). Differentiation of M in this case yields

$$\partial_2 M(a,r) \equiv \frac{\partial M}{\partial r}(a,r) = -2\pi a\int_0^\infty \rho\breve{w}(\rho)J_1(r\rho)J_1(a\rho)d\rho \tag{5.8}$$

implying that

$$\mathbf{sgn}\big(\partial_2 M(a,r)\big) = \mathbf{sgn}\big(\partial_2 M(r,a)\big).$$

Consequently $\frac{\partial M}{\partial r}(a,r) < 0$ also for $r < a$. Hence U is monotonically decreasing in r for any monotonic synaptic weight function w.

Exponential weight function. Consider the radially symmetric exponential weight function and its Hankel representation

$$w(r) = \frac{1}{2\pi}e^{-r}, \qquad \breve{w}(r) = \frac{1}{2\pi}\frac{1}{(1+\rho^2)^{\frac{3}{2}}}. \tag{5.9}$$

The condition for the existence of a stationary pulse is then given by

$$(1+\beta)\kappa = \mathbf{M}(a) + I(a) \equiv G(a), \tag{5.10}$$

where

$$\mathbf{M}(a) \equiv M(a,a) = a\int_0^\infty \frac{1}{(\rho^2+1)^{\frac{3}{2}}} J_0(a\rho)J_1(a\rho)d\rho. \tag{5.11}$$

The function $G(a)$ is plotted in Figure 8 for a range of input amplitudes $\mathcal{I}$, with horizontal lines indicating different values of $\hat{\kappa}$; intersection points determine the existence of stationary pulse solutions. Note that the integral expression on the right-hand side of (5.11) can be evaluated explicitly in terms of finite sums of modified Bessel and Struve functions; see Appendix B.

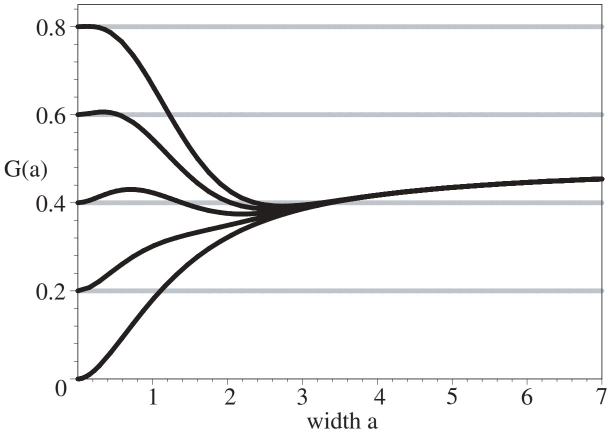


Figure 8. *Plot of $G(a)$ defined in (5.10) as a function of pulse width a for various values of input amplitude $\mathcal{I}$ and for fixed input width $\sigma = 1$.*

We proceed in the same fashion as in the one-dimensional case and generate stationary-pulse existence curves for the exponential weight function. Qualitatively the catalogue of bifurcation scenarios is similar, although there is now an additional case. In one dimension we have $G'(0) > 0$ so that there are always at least two solution branches when $\hat{\kappa} > 1/2$. On the other hand, in two dimensions we have $G'(0) < 0$ for sufficiently large input amplitude $\mathcal{I}$ so that it is possible to find only one solution branch for large $\hat{\kappa}$, that is, when $\hat{\kappa} > \kappa_0$ for some critical value $\kappa_0 > 1/2$. Hence, there are three distinct cases as shown in Figure 9. The effect of varying σ identically follows the one-dimensional case.

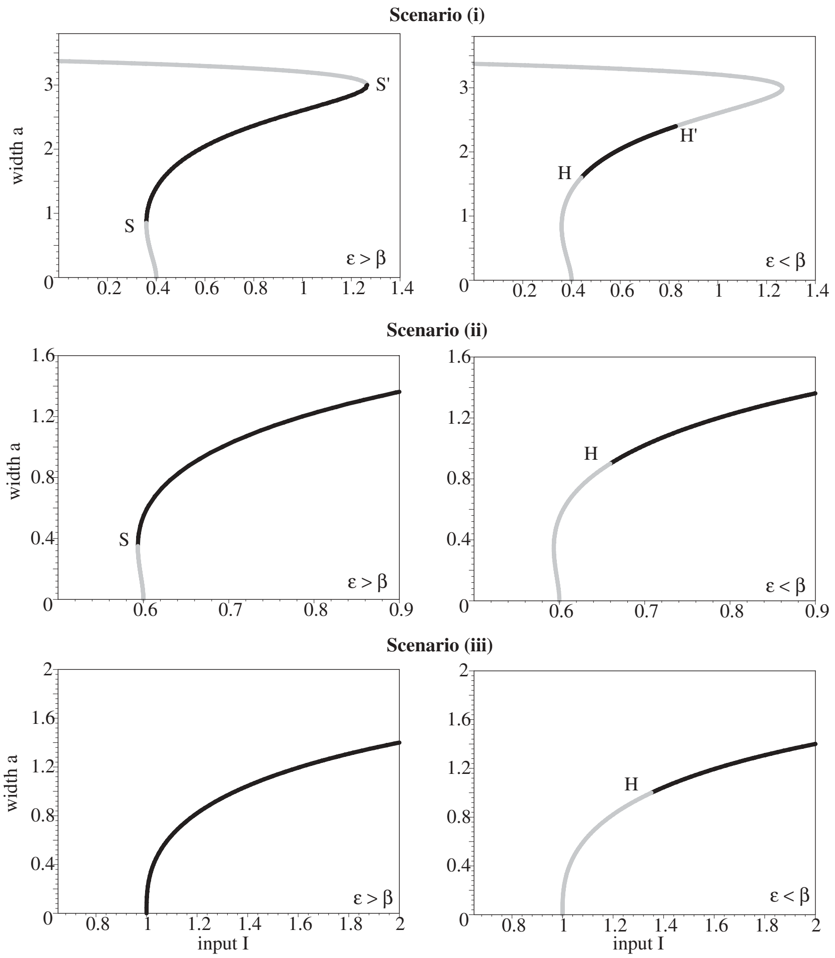


Figure 9. *Two-dimensional stationary-pulse existence curves for an exponential weight distribution:* (i) $\kappa_c < \hat{\kappa} < \frac{1}{2}$, (ii) $\frac{1}{2} < \hat{\kappa} < \kappa_0$, *and* (iii) $\kappa_0 < \hat{\kappa}$. *Other parameter values are* $\beta = 1$, $\sigma = 1.0$. *Black indicates stability, whereas gray indicates instability of the stationary pulse. Saddle-node bifurcation points are indicated by* S, S' *and Hopf bifurcation points by* H, H'.

Modified Bessel weight function. In the case of the exponential weight function w we do not have a closed form for the integral in (5.6). Here we consider a nearby problem where we are able to construct the stationary-pulse solution explicitly. Consider the radially symmetric weight function, normalized to unity,

$$w(r) = \frac{2}{3\pi}\big(K_0(r) - K_0(2r)\big), \tag{5.12}$$

where K_ν is the modified Bessel function of the second kind, whose Hankel transform is

$$\breve{w}(r) = \frac{2}{3\pi}\left(\frac{1}{\rho^2+1} - \frac{1}{\rho^2+2^2}\right). \tag{5.13}$$

The coefficient $2/3\pi$ is chosen so that there is a good fit with the exponential distribution as shown in Figure 10(a). Note that

$$w(0) = \frac{1}{2\pi}\frac{4\ln(2)}{3} \approx \frac{1}{2\pi}(0.924), \qquad w(r) \sim \frac{1}{3}\left(\frac{\sqrt{2}e^{-r} - e^{-2r}}{\sqrt{r}}\right) \quad \text{for large } r.$$

Substituting (5.13) into (5.3), we can explicitly compute the resulting integral using Bessel functions:

$$a\int_0^\infty \frac{1}{\rho^2+s^2} J_0(r\rho)J_1(a\rho)d\rho = \begin{cases} \frac{a}{s}I_1(sa)K_0(sr) & \text{for } r \geq a, \\ \frac{1}{s^2} - \frac{a}{s}I_0(sr)K_1(sa) & \text{for } r < a, \end{cases}$$

where I_ν is the modified Bessel function of the first kind. Substituting into (5.3) shows that

$$\begin{aligned} M(a,r) &= 2\pi a \int_0^\infty \breve{w}(\rho)J_0(r\rho)J_1(a\rho)d\rho \\ &= \frac{4}{3}a\int_0^\infty \left(\frac{1}{\rho^2+1} - \frac{1}{\rho^2+2^2}\right) J_0(r\rho)J_1(a\rho)d\rho \\ &= \begin{cases} \frac{4}{3}\left(aI_1(a)K_0(r) - \frac{a}{2}I_1(2a)K_0(2r)\right) & \text{for } r \geq a, \\ 1 - \frac{4}{3}\left(aI_0(r)K_1(a) + \frac{a}{2}I_0(2r)K_1(2a)\right) & \text{for } r < a. \end{cases} \end{aligned}$$

The condition for the existence of a stationary pulse of radius a is thus given by (5.10) with

$$\mathbf{M}(a) = \frac{4}{3}\left(aI_1(a)K_0(a) - \frac{a}{2}I_1(2a)K_0(2a)\right). \tag{5.14}$$

An example of an exact pulse solution is shown in Figure 10(b).

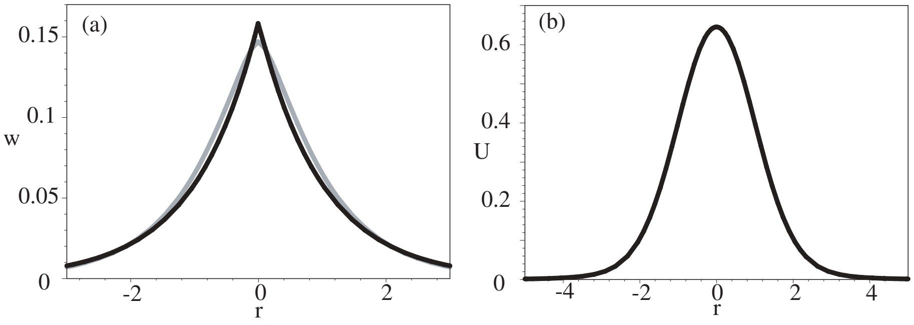


Figure 10. (a) *Synaptic weight functions, exponential weight in black and modified Bessel weight in gray.* (b) *Stationary-pulse solution with half-width,* $a = 1$, *generated by the modified Bessel weight function with* $\kappa = 0.4$, $\beta = 1$, $\mathcal{I} = 1$.

5.2. Stability analysis. We now analyze the evolution of small time-dependent perturbations of the stationary-pulse solution through linear stability analysis. We investigate saddle-node and Hopf bifurcations of the stationary pulse by relating the eigenvalues to the gradient of the Gaussian input I. The behavior of the system near and beyond the Hopf bifurcation is then studied numerically as in one dimension.

Spectral analysis of the linearized operator. Equation (5.1) is linearized about the stationary solution (U, Q) by introducing the time-dependent perturbations

$$\begin{aligned} u(\mathbf{r}, t) &= U(r) + \varphi(\mathbf{r}, t), \\ q(\mathbf{r}, t) &= Q(r) + \psi(\mathbf{r}, t) \end{aligned}$$

with $Q = U$ and expanding to first-order in φ, ψ. This leads to the linearized system of equations

$$\begin{aligned} \frac{\partial \varphi}{\partial t}(\mathbf{r}, t) &= -\varphi(\mathbf{r}, t) + \int_{\mathbb{R}^2} w(|\mathbf{r} - \mathbf{r}'|) H'(U(r') - \kappa)\varphi(\mathbf{r}', t) d\mathbf{r}' - \beta\psi(\mathbf{r}, t), \\ \frac{1}{\epsilon}\frac{\partial \psi}{\partial t}(\mathbf{r}, t) &= -\psi(\mathbf{r}, t) + \varphi(\mathbf{r}, t). \end{aligned}$$

We separate variables

$$\begin{aligned} \varphi(\mathbf{r}, t) &= \varphi(\mathbf{r}) e^{\lambda t}, \\ \psi(\mathbf{r}, t) &= \psi(\mathbf{r}) e^{\lambda t} \end{aligned}$$

to obtain the system

$$\begin{aligned} \lambda\varphi(\mathbf{r}) &= -\varphi(\mathbf{r}) + \int_{\mathbb{R}^2} w(|\mathbf{r} - \mathbf{r}'|) H'(U(r') - \kappa)\varphi(\mathbf{r}') d\mathbf{r}' - \beta\psi(\mathbf{r}), \\ \frac{\lambda}{\epsilon}\psi(\mathbf{r}) &= -\psi(\mathbf{r}) + \varphi(\mathbf{r}). \end{aligned} \tag{5.15}$$

Solving (5.15), we find

$$\left(\lambda + 1 + \frac{\beta\epsilon}{\lambda + \epsilon}\right)\varphi(\mathbf{r}) = \int_{\mathbb{R}^2} w(|\mathbf{r} - \mathbf{r}'|) H'(U(r') - \kappa)\varphi(\mathbf{r}') d\mathbf{r}'. \tag{5.16}$$

Introducing polar coordinates $\mathbf{r} = (r, \theta)$ and using the result

$$H'(U(r) - \kappa) = \delta(U(r) - \kappa) = \frac{\delta(r - a)}{|U'(a)|},$$

we obtain

$$\begin{aligned} \left(\lambda + 1 + \frac{\epsilon\beta}{\lambda + \epsilon}\right)\varphi(\mathbf{r}) &= \int_0^{2\pi}\int_0^{\infty} w(|\mathbf{r} - \mathbf{r}'|)\frac{\delta(r' - a)}{|U'(a)|}\varphi(\mathbf{r}') r' dr' d\theta' \\ &= \frac{a}{|U'(a)|}\int_0^{2\pi} w(|\mathbf{r} - \mathbf{a}'|)\varphi(a, \theta') d\theta', \end{aligned} \tag{5.17}$$

where $\mathbf{a}' = (a, \theta')$

We consider the following two cases. (i) The function φ satisfies the condition

$$\int_0^{2\pi} w(|\mathbf{r} - \mathbf{a}'|)\varphi(a, \theta') d\theta' = 0$$

for all $\mathbf{r}$. The integral equation reduces to

$$\lambda + 1 + \frac{\beta\epsilon}{\lambda+\epsilon} = 0,$$

yielding the eigenvalues

$$\lambda_{\pm}^{(0)} = \frac{-(1+\epsilon) \pm \sqrt{(1+\epsilon)^2 - 4\epsilon(1+\beta)}}{2}.$$

This is part of the essential spectrum and is identical to the one-dimensional case: it is negative and does not cause instability. (ii) φ does not satisfy the above condition, and we must study the solutions of the integral equation

$$\mu\varphi(r,\theta) = a\int_0^{2\pi} \mathcal{W}(a,r;\theta-\theta')\varphi(a,\theta')d\theta',$$

where

$$\lambda + 1 + \frac{\epsilon\beta}{\lambda+\epsilon} = \frac{\mu}{|U'(a)|} \tag{5.18}$$

and $\mathcal{W}(a,r;\phi) = w\left(\sqrt{r^2+a^2-2ra\cos\phi}\right)$. This equation demonstrates that $\varphi(r,\theta)$ is determined completely by its values $\varphi(a,\theta)$ on the restricted domain $r=a$. Hence we need only consider $r=a$, yielding the integral equation

$$\mu\varphi(a,\theta) = a\int_0^{2\pi} \mathcal{W}(a,a;\phi)\varphi(a,\theta-\phi)d\phi. \tag{5.19}$$

The solutions of this equation are exponential functions $e^{\gamma\theta}$, where γ satisfies

$$a\int_0^{2\pi} \mathcal{W}(a,a;\phi)e^{-\gamma\phi}d\phi = \mu.$$

By the requirement that φ is 2π-periodic in θ, it follows that $\gamma = in$, where $n\in\mathbb{Z}$. Thus the integral operator with kernel $\mathcal{W}$ has a discrete spectrum given by

$$\begin{aligned}\mu_n &= a\int_0^{2\pi} \mathcal{W}(a,a;\phi)e^{-in\pi\phi}d\phi \\ &= a\int_0^{2\pi} w\left(\sqrt{a^2+a^2-2a^2\cos\phi}\right)e^{-in\phi}d\phi \\ &= 2a\int_0^{\pi} w\left(2a\sin(\phi)\right)e^{-2in\phi}d\phi \end{aligned} \tag{5.20}$$

(after rescaling ϕ). Note that μ_n is real since

$$\mathrm{Im}\{\mu_n(a)\} = -2a\int_0^{\pi} w(2a\sin(\phi))\sin(2n\phi)d\phi = 0;$$

i.e., the integrand is odd-symmetric about $\pi/2$. Hence,

$$\mu_n(a) = \mathrm{Re}\{\mu_n(a)\} = 2a\int_0^\pi w(2a\sin(\phi))\cos(2n\phi)d\phi$$

with the integrand even-symmetric about $\frac{\pi}{2}$. Since $w(r)$ is a positive function of r, it follows that

$$\mu_n(a) \leq 2a\int_0^\pi w(2a\sin(\phi))\,|\cos(2n\phi)|d\phi \leq 2a\int_0^\pi w(2a\sin(\phi))d\phi = \mu_0(a).$$

Given the set of solutions $\{\mu_n(a),\, n\geq 0\}$ for a pulse of width a (assuming that it exists), we obtain from (5.18) the following pairs of eigenvalues:

$$\lambda_n^{\pm} = \frac{1}{2}\left(-\Lambda_n \pm \sqrt{\Lambda_n^2 - 4\epsilon(1+\beta)(1-\Gamma_n)}\right), \tag{5.21}$$

where

$$\Lambda_n = 1+\epsilon-\Gamma_n(1+\beta), \qquad \Gamma_n = \frac{\mu_n(a)}{|U'(a)|(1+\beta)}. \tag{5.22}$$

Stability of the two-dimensional pulse requires that

$$\Lambda_n > 0, \qquad \Gamma_n < 1 \qquad \text{for all } n\geq 0.$$

This reduces to the stability conditions

$$\begin{aligned} \epsilon > \beta: &\qquad \Gamma_n < 1 &&\qquad \text{for all } n \geq 0,\\ \epsilon < \beta: &\qquad \Gamma_n < \frac{1+\epsilon}{1+\beta} &&\qquad \text{for all } n\geq 0. \end{aligned} \tag{5.23}$$

Next we rewrite the stability conditions in terms of the gradient of the input $D(a) = |I'(a)|$. From (5.2), (5.8), and (5.9) we have

$$U'(a) = \frac{1}{1+\beta}\big(-M_r(a) + I'(a)\big),$$

where

$$M_r(a) \equiv -\frac{\partial}{\partial r}M(a,r)\bigg|_{r=a} = 2\pi a\int_0^\infty \rho\breve{w}(\rho)J_1(a\rho)J_1(a\rho)d\rho. \tag{5.24}$$

We have already established in section 5.1 that $M_r(a) > 0$. Hence,

$$\begin{aligned} |\,U'(a)\,| &= \left(\frac{1}{1+\beta}\right)\big|-M_r(a)+I'(a)\big| \\ &= \left(\frac{1}{1+\beta}\right)\big(M_r(a)+D(a)\big). \end{aligned} \tag{5.25}$$

The stability conditions (5.23) thus become

$$\begin{aligned} \epsilon > \beta: &\qquad D(a) > \mu_n(a) - M_r(a) &&\qquad \text{for all } n \geq 0, \\ \epsilon < \beta: &\qquad D(a) > \left(\frac{1+\beta}{1+\epsilon}\right)\mu_n(a) - M_r(a) &&\qquad \text{for all } n \geq 0. \end{aligned}$$

Finally, using the fact that $\mu_0(a) \geq \mu_n(a)$ for all $n \geq 1$ and $M_r(a) > 0$, we obtain the reduced stability conditions

$$\epsilon > \beta: \qquad D(a) > \mu_0(a) - M_r(a) \equiv D_{\mathbf{SN}}(a), \tag{5.26}$$

$$\epsilon < \beta: \qquad D(a) > \left(\frac{1+\beta}{1+\epsilon}\right)\mu_0(a) - M_r(a) \equiv D_c(a). \tag{5.27}$$

We now relate stability of the stationary pulse to the gradient D on different branches of the existence curves shown in Figure 9 for $w(r)$ given by the exponential distribution (5.9). In this case the integral expressions for $\mu_0(a)$, $M_r(a)$, and $\mathbf{M}(a)$ can be evaluated explicitly in terms of finite sums of modified Bessel and Struve functions; see Appendix B.

Stability for $\epsilon > \beta$. Equation (5.10) implies that $G'(a) = \mathbf{M}'(a) + I'(a)$. Since $\mathbf{M}(a) = M(a, a)$, it follows from (5.3), (5.20), and (5.24) that

$$\begin{aligned} \mathbf{M}'(a) &= \partial_1 M(a,a) + \partial_2 M(a,a) \\ &= \mu_0(a) - M_r(a), \end{aligned} \tag{5.28}$$

and hence

$$G'(a) = \mu_0(a) - M_r(a) - D(a). \tag{5.29}$$

We now see that the stability condition (5.26) is satisfied when $G'(a) < 0$ and is not satisfied when $G'(a) > 0$. Saddle-node bifurcations occur when $G'(a)$, that is, when $D(a)$ passes through $D_{\mathbf{SN}}(a) = \mu_0(a) - M_r(a)$. This establishes the stability of the middle branch in case (i) and the upper branch of cases (ii) and (iii) shown in the left-hand column of Figure 9.

Hopf curves for $\epsilon < \beta$. A Hopf bifurcation occurs when $\Lambda_0 = 0$ and $\Gamma_0 < 1$, or equivalently, when $D(a) = D_c(a)$. Since $\mu_0(a) > 0$ it follows from (5.26) and (5.27) that $D_c(a) > D_{\mathbf{SN}}(a)$, and hence Hopf bifurcations occur only on the branches that are stable when $\epsilon > \beta$. As in the one-dimensional case, the Hopf and saddle-node points coincide when $\epsilon = \beta$, and so we expect, as ϵ decreases from β, the Hopf bifurcation point(s) to traverse these previously stable branches from the saddle-node point(s). Again, in order to show this more explicitly, we find a relationship for $D(a)$ that is independent of the input amplitude $\mathcal{I}$. Using (5.10), the input gradient D can be related as

$$\begin{aligned} D(a) &= |I'(a)| \\ &= \frac{a}{\sigma^2} I(a) \\ &= \frac{a}{\sigma^2}\left(\kappa(1+\beta) - \mathbf{M}(a)\right). \end{aligned} \tag{5.30}$$

For each of the cases discussed in section 3.1, we examine graphically the crossings of the curves $D(a), D_c(a)$: stability corresponds to $D(a) > D_c(a)$ with Hopf points at $D(a) = D_c(a)$.

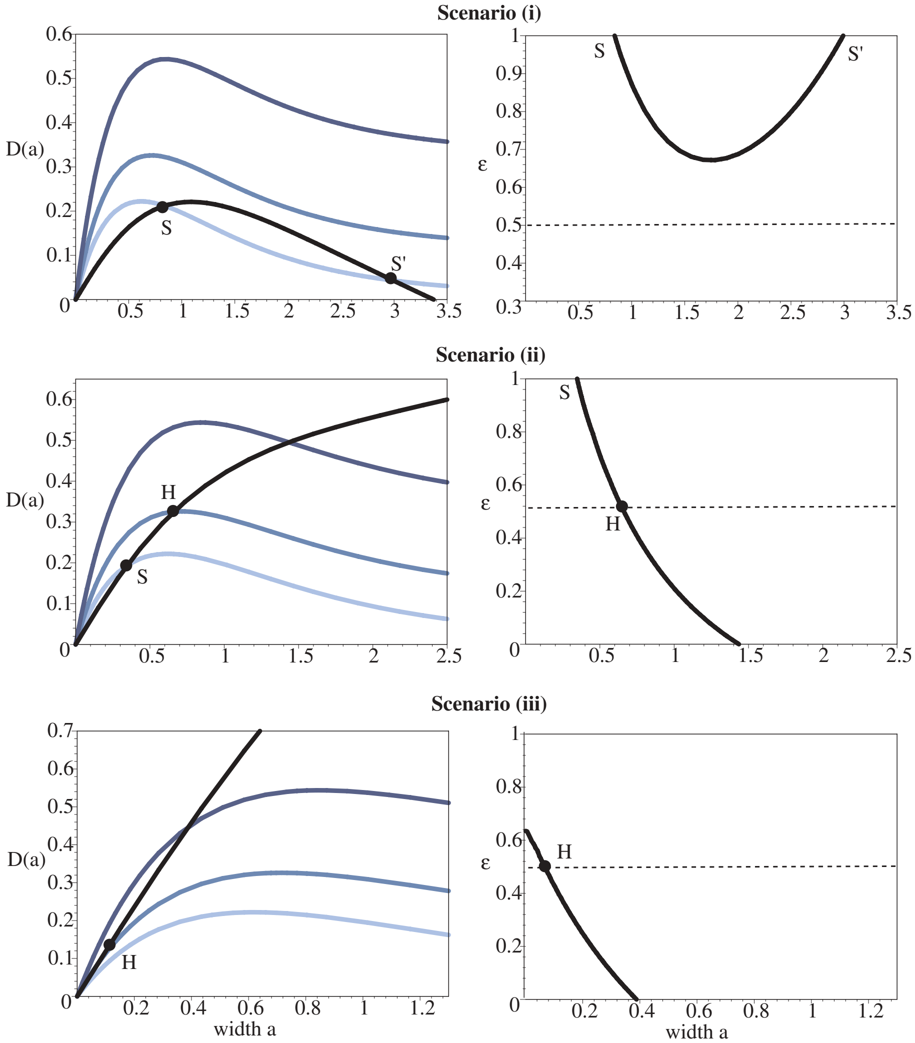


Figure 11. *Left column: Gradient curves for the various bifurcation scenarios shown in Figure* 9: (i) $\kappa_c < \hat{\kappa} < \frac{1}{2}$, (ii) $\frac{1}{2} < \hat{\kappa} < \kappa_0$, *and* (iii) $\kappa_0 < \hat{\kappa}$. *The thick solid curve shows the input gradient* $D(a)$ *as a function of pulse width* a*. The lighter curves show the critical gradient* $D_c(a)$ *as function of* a *for* $\epsilon = 0.0, 0.5, 1.0$ *and* $\beta = 1$*. For a given value of* $\epsilon < \beta$*, a stationary pulse of width* a *is stable provided that* $D(a) > D_c(a)$*. A pulse loses stability via a Hopf bifurcation at the intersection points* $D(a) = D_c(a)$*. The Hopf bifurcation points for* $\epsilon = 0.5$ *are indicated by* H*; in the first scenario there are no Hopf points at this particular value of* ϵ*. In the limit* $\epsilon \to \beta$*, we have* $H \to S$*. Right column: Corresponding Hopf stability curves in the* (a, ϵ)*-plane.*

The results are displayed in Figure 11. Note that as in one dimension, sufficiently wide pulse solutions are always stable, as can be established by studying the asymptotic behavior of $D(a)$ and $D_c(a)$; see Appendix B.

Two-dimensional breathers. Analogous to the one-dimensional case, we find numerically that the upper branch in scenarios (ii) and (iii) can undergo a supercritical Hopf bifurcation leading to the formation of a two-dimensional breather. An example is shown in Figure 12, which was obtained using a Runge–Kutta scheme on a 300×300 grid with a time step of 0.02.

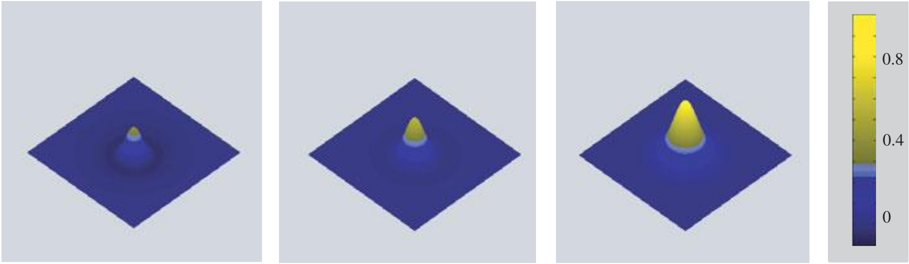


Figure 12. *Two-dimensional breather with $\beta = 4$, $\kappa = 0.25$, $\epsilon = 0.1$, $\mathcal{I} = 0.26$. Clicking on the above image displays the associated movie.*

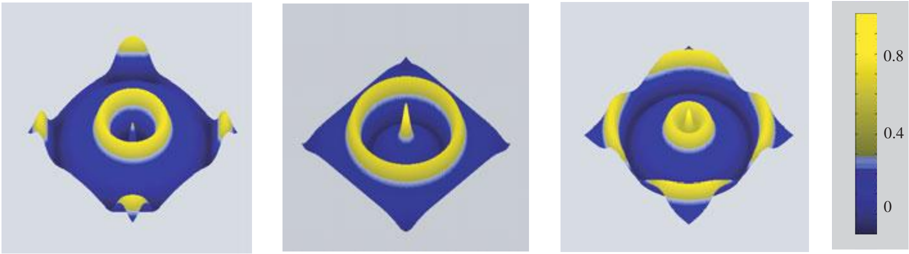


Figure 13. *Two-dimensional pulse emitter with $\beta = 4$, $\kappa = 0.2$, $\epsilon = 0.1$, $\mathcal{I} = 0.2$. Clicking on the above image displays the associated movie.*

Moreover, the breather can undergo a secondary instability, resulting in the periodic emission of circular target waves; see Figure 13.

6. Discussion.

In this paper we have shown that a localized external input can induce oscillatory behavior in an excitatory neural network in the form of breathing pulses and that these breathers can subsequently act as sources of wave emission. Interestingly, following some initial excitation, breathers can be supported by subthreshold inputs. From a mathematical perspective, there are a number of directions for future work. First, one could try to develop some form of weakly nonlinear analysis in order to determine analytically whether or not the Hopf instability of the stationary pulse is supercritical or subcritical. It would also be interesting to explore more fully the behavior around the degenerate bifurcation point $\epsilon = \beta$, where there exists a pair of zero eigenvalues of the associated linear operator that is suggestive of a Takens–Bogdanov bifurcation. The latter would predict that for certain parameter values around the degenerate bifurcation point, the periodic orbit arising from the Hopf bifurcation could be annihilated in a homoclinic bifurcation associated with another unstable stationary pulse. This could provide one mechanism for the disappearance of the breather as the amplitude of the external input is reduced. (A pulse-emitter does not occur when $\epsilon \approx \beta$.) Another extension is to consider a smooth nonlinear output function $f(u) = 1/(1 + \mathrm{e}^{-\gamma(u-\kappa)})$, which reduces to the Heaviside function in the high gain limit $\gamma \to \infty$. In the case of sufficiently slow adaptation (small ϵ), it might be possible to use singular perturbation methods along the lines of Pinto and Ermentrout [25], who established the existence of traveling pulses in a homogeneous network with smooth f. Finally, it would be interesting to extend our analysis to the case of traveling waves locking to a moving stimulus; the associated stability analysis would then involve Evans functions. From an experimental perspective, our results could be tested by introducing an inhomogeneous current into a cortical slice and searching for these oscillations. One potential difficulty of such an experiment is that persistent currents tend to burn out neurons. An alternative approach might be to use some form of pharmacological manipulation of NMDA receptors, for example, or an application of an external electric field that modifies the effective threshold of the neurons. Note that the usual method for inducing traveling waves in cortical slices (and in corresponding computational models) is to introduce short-lived current injections; once the wave is formed it propagates in a homogeneous medium.

Appendix A. We used the following biophysical model functions and parameter values in section 4.2:

$$\begin{aligned} V_{\text{syn}} &= -45 \text{ mV}, & g_{\text{syn}} &= 20 \text{ mS/cm}^2, \\ V_{\text{K}} &= -100 \text{ mV}, & g_{\text{K}} &= 80 \text{ mS/cm}^2, \\ V_{\text{Na}} &= 50 \text{ mV}, & g_{\text{Na}} &= 100 \text{ mS/cm}^2, \\ V_{\text{L}} &= -67 \text{ mV}, & g_{\text{L}} &= 0.2 \text{ mS/cm}^2, \\ F &= 1\mu\text{F/cm}^2, & g_q &= 3 \text{ mS/cm}^2, \end{aligned}$$

$$\begin{aligned} \alpha_m(v) &= 0.32(54+v)/(1-\exp(-(v+54)/4)), \\ \beta_m(v) &= 0.28(v+27)/(\exp((v+27)/5)-1), \\ \alpha_h(v) &= 0.128\exp(-(50+v)/18), \\ \beta_h(v) &= 4/(1+\exp(-(v+27)/5)), \\ \alpha_n(v) &= 0.032(v+52)/(1-\exp(-(v+52)/5)), \\ \beta_n(v) &= 0.5\exp(-(57+v)/40), \end{aligned}$$

where

$$p_\infty(v) = \frac{\alpha_p(v)}{\alpha_p(v)+\beta_p(v)}, \qquad \tau_p(v) = \frac{1}{\alpha_p(v)+\beta_p(v)}, \qquad p \in \{m,n,h\},$$

$$q_\infty(v) = \frac{1}{1+e^{(-(v+35)/20)}}, \qquad \tau_q(v) = \frac{1000}{3.3e^{(v+35)/20}+e^{-(v+35)/20}},$$

$$\tau = 1, \qquad K(V) = \frac{1}{1+e^{-(V+50)}}.$$

Appendix B. In this appendix we evaluate the gradient functions $D(a)$ and $D_c(a)$ of (5.30) and (5.27) for the exponential weight distribution (5.9). We then determine their asymptotic behavior for large pulse width a, thus establishing the stability of stationary pulses in the limit $a \to \infty$. First, from (5.20), we can write

$$\begin{aligned}\mu_0(a) &= \frac{a}{\pi}\int_0^{\pi} \exp(-2a\sin\phi)d\phi \\ &= \frac{2a}{\pi}\int_0^{\frac{\pi}{2}} \exp(-2a\cos\phi)d\phi \\ &= a\left(\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\cosh(-2a\cos\phi)d\phi\right) - a\left(\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\sinh(-2a\cos\phi)d\phi\right) \qquad \text{(B.1)} \\ &= a\left(I_0(2a) - \mathbf{L}_0(2a)\right), \qquad \text{(B.2)}\end{aligned}$$

where I_ν is a modified Bessel function and $\mathbf{L}_\nu$ denotes a modified Struve function [33]. Second, from (5.24),

$$\begin{aligned}M_r(a) &= a\int_0^{\infty} \frac{\rho}{(\rho^2+1)^{\frac{3}{2}}} J_1(a\rho)J_1(a\rho) \\ &= a\big(\mathbf{L}_0(2a) - I_0(2a)\big) - \big(\mathbf{L}_1(2a) - I_1(2a)\big), \qquad \text{(B.3)}\end{aligned}$$

where

$$I_1(2a) = \frac{4a}{\pi}\int_0^{\pi/2} \cosh(2a\cos\theta)\sin^2\theta d\theta$$

and

$$\mathbf{L}_1(2a) = \frac{4a}{\pi}\int_0^{\pi/2} \sinh(2a\cos\theta)\sin^2\theta d\theta.$$

Equation (5.27) then implies that

$$D_c(a) = \frac{\beta+\epsilon+2}{1+\epsilon} a\big(I_0(2a) - \mathbf{L}_0(2a)\big) - \big(I_1(2a) - \mathbf{L}_1(2a)\big). \qquad \text{(B.4)}$$

Using the asymptotic expansions for large a,

$$\begin{aligned}I_0(2a) - \mathbf{L}_0(2a) &\sim \frac{1}{\pi}\left(\frac{1}{a} + \frac{1}{4a^2}\right), \\ I_1(2a) - \mathbf{L}_1(2a) &\sim \frac{2}{\pi}\left(1 - \frac{1}{4a^2}\right), \qquad \text{(B.5)}\end{aligned}$$

we deduce that

$$D_c(a) \sim \frac{1}{\pi}\left(\frac{\beta-\epsilon}{1+\epsilon}\right) + \left(\frac{1}{4}\left(\frac{\beta-\epsilon}{1+\epsilon}\right) + 1\right)\frac{1}{\pi a^2}. \qquad \text{(B.6)}$$

Similarly, from (5.11) we have

$$\begin{aligned}\mathbf{M}(a) &= a\int_0^{\infty} \frac{1}{(\rho^2+1)^{\frac{3}{2}}} J_0(a\rho)J_1(a\rho)d\rho \\ &= \left(\frac{1}{2} + aI_1(2a) - \frac{1}{2}I_0(2a)\right) - \left(\frac{2a}{\pi} + a\mathbf{L}_1(2a) - \frac{1}{2}\mathbf{L}_0(2a)\right). \qquad \text{(B.7)}\end{aligned}$$

Equation (5.30) and the asymptotic expansions (B.5) then imply that

$$D(a) \sim \frac{a}{\sigma^2}\left((1+\beta)\kappa - \frac{1}{2}\right) + \frac{1}{\pi\sigma^2} - \frac{1}{8\pi a^2}. \tag{B.8}$$

Finally, combining (B.6) and (B.8),

$$D(a) - D_c(a) \sim \frac{a}{\sigma^2}\left((1+\beta)\kappa - \frac{1}{2}\right) + \frac{1}{\pi\sigma^2} - \left(\frac{\beta-\epsilon}{1+\epsilon}\right) + \mathcal{O}(a^{-2}). \tag{B.9}$$

From this we conclude that for all $\sigma > 0$, a stationary-pulse solution (if it exists) is stable in the limit $a \to \infty$, provided that $\hat{\kappa} > 1/2$.

REFERENCES

[1] S. Amari, *Dynamics of pattern formation in lateral inhibition type neural fields*, Biol. Cybernetics, 27 (1977), pp. 77–87.

[2] M. Bode, *Front-bifurcations in reaction-diffusion systems with inhomogeneous parameter distributions*, Phys. D, 106 (1997), pp. 270–286.

[3] 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.

[4] P. C. Bressloff, *Traveling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[5] P. C. Bressloff, S. E. Folias, A. Pratt, and Y.-X. Li, *Oscillatory waves in inhomogeneous neural media*, Phys. Rev. Lett., 91 (2003), 178101.

[6] P. C. Bressloff and S. E. Folias, *Front bifurcations in an excitatory neural network*, SIAM J. Appl. Math., 65 (2004), pp. 131–151.

[7] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Propagation of excitation in neural network models*, J. Neurophysiol., 60 (1988), pp. 1695–1713.

[8] G. B. Ermentrout and J. B. McLeod, *Existence and uniqueness of travelling waves for a neural network*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.

[9] G. B. Ermentrout, *The analysis of synaptically generated traveling waves*, J. Comput. Neurosci., 5 (1998), pp. 191–208.

[10] G. B. Ermentrout, *Neural networks as spatial pattern forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[11] G. B. Ermentrout and D. Kleinfeld, *Traveling electrical waves in cortex: Insights from phase dynamics and speculation on a computational role*, Neuron, 29 (2001), pp. 33–44.

[12] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiol., 78 (1997), pp. 1199–1211.

[13] A. Hagberg and E. Meron, *Pattern formation in nongradient reaction-diffusion systems: The effects of front bifurcations*, Nonlinearity, 7 (1994), pp. 805–835.

[14] A. Hagberg, E. Meron, I. Rubinstein, and B. Zaltzman, *Controlling domain patterns far from equilibrium*, Phys. Rev. Lett., 76 (1996), pp. 427–430.

[15] M. A. P. Idiart and L. F. Abbott, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.

[16] J. P. Keener, *Propagation of waves in an excitable medium with discrete release sites*, SIAM J. Appl. Math., 61 (2000), pp. 317–334.

[17] J. P. Keener, *Homogenization and propagation in the bistable equation*, Phys. D, 136 (2000), pp. 1–17.

[18] D. Kleinfeld, K. R. Delaney, M. S. Fee, J. A. Flores, D. W. Tank, and A. Galperin, *Dynamics of propagating waves in the olfactory network of a terrestrial mollusk: An electrical and optical study*, J. Neurophysiol., 72 (1994), pp. 1402–1419.

[19] 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.

[20] C. R. Laing and W. C. Troy, *PDE methods for nonlocal models*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 487–516.

[21] Y. W. Lam, L. B. Cohen, M. Wachowiak, and M. R. Zochowski, *Odors elicit three different oscillations in the turtle olfactory bulb*, J. Neurosci., 20 (2000), pp. 749–762.

[22] Y.-X. Li, *Tango waves in a bidomain model of fertilization calcium waves*, Phys. D, 186 (2003), pp. 27–49.

[23] R. Malach, Y. Amir, M. Harel, and A. Grinvald, *Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 10469–10473.

[24] M. A. L. Nicolelis, L. A. Baccala, R. C. S. Lin, and J. K. Chapin, *Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system*, Science, 268 (1995), pp. 1353–1358.

[25] 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.

[26] A. Prat and Y.-X. Li, *Stability of front solutions in inhomogeneous media*, Phys. D, 186 (2003), pp. 50–68.

[27] J. C. Prechtl, L. B. Cohen, B. Pasaram, P. P. Mitra, and D. Kleinfeld, *Visual stimuli induce waves of electrical activity in turtle cortex*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 7621–7626.

[28] J. Rinzel and D. Terman, *Propagation phenomena in a bistable reaction-diffusion system*, SIAM J. Appl. Math., 42 (1982), pp. 1111–1137.

[29] P. R. Roelfsema, A. K. Engel, P. Konig, and W. Singer, *Visuomotor integration is associated with zero time-lag synchronization among cortical areas*, Nature, 385 (1997), pp. 157–161.

[30] J. E. Rubin and W. C. Troy, *Sustained spatial patterns of activity in neuronal populations without recurrent excitation*, SIAM J. Appl. Math., 64 (2004), pp. 1609–1635.

[31] P. Schutz, M. Bode, and H.-G. Purwins, *Bifurcations of front dynamics in a reaction-diffusion system with spatial inhomogeneities*, Phys. D, 82 (1995), pp. 382–397.

[32] J. G. Taylor, *Neural "bubble" dynamics in two dimensions: Foundations*, Biol. Cybernetics, 80 (1999), pp. 303–409.

[33] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, UK, 1952.

[34] H. Werner and T. Richter, *Circular stationary solutions in two-dimensional neural fields*, Biol. Cybernetics, 85 (2001), pp. 211–217.

[35] J.-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neurosci., 19 (1999), pp. 5005–5015.

[36] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.

[37] T. Yoshioka, G. G. Blasdel, J. B. Levitt, and J. S. Lund, *Relation between patterns of intrinsic lateral connectivity, ocular dominance, and cytochrome oxidase-reactive regions in macaque monkey striate cortex*, Cerebral Cortex, 6 (1996), pp. 297–310.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 409–432

First-Order Averaging Principles for Maps with Applications to Accelerator Beam Dynamics[*]

H. Scott Dumas[†], James A. Ellison[‡], and Mathias Vogt[§]

Abstract. For slowly evolving, discrete-time–dependent systems of difference equations (iterated maps), we believe that the simplest means of demonstrating the validity of the averaging method at first order is by way of a lemma that we call the Besjes inequality. In this paper, we develop the Besjes inequality for identity maps with perturbations that are (i) at low-order resonance (periodic with short period) and (ii) far from low-order resonance in discrete time. We use these inequalities to prove corresponding first-order averaging principles, together with a principle of adiabatic invariance on extended timescales, and we generalize and apply these mathematical results to model problems in accelerator beam dynamics and to the Hénon map.

Key words. averaging method, averaging principle, difference equations, iterated maps, small divisors, adiabatic invariance, accelerator beam dynamics, kick-rotate model, Hénon map

AMS subject classifications. 39A11, 37J40, 70K65, 70H11, 78A35, 70Fxx

DOI. 10.1137/030600436

1. Introduction.

In broadest terms, the method of averaging (or "averaging principle") may be described as follows: to approximate the evolution of a system with motions occurring on both fast and slow timescales, one uses a simpler system (often loosely called a "normal form") obtained by somehow averaging over the fast motion of the original system. In the context of difference equations (or "iterated maps"), the most elementary situation to which the method applies occurs in periodic systems of the form

$$x_{n+1} = x_n + \varepsilon f(x_n, n), \tag{1.1}$$

where $x_n \in U \subseteq \mathbf{R}^d$, $n \in \mathbf{N}$, $\varepsilon > 0$ is a small parameter, and $f : U \times \mathbf{N} \to \mathbf{R}^d$ is a bounded, locally x-Lipschitz, discrete-time–dependent function of period p in n. Solutions of system (1.1) are approximated by solutions of the associated averaged system

$$y_{n+1} = y_n + \varepsilon \widehat{f}(y_n), \tag{1.2}$$

where the autonomous function $\widehat{f} : U \to \mathbf{R}^d$ (the *average of* f) is given by $\widehat{f}(y) := (1/p)\sum_{n=0}^{p-1} f(y, n)$. In this context the averaging principle asserts that solutions x_n of (1.1) and y_n of (1.2) that start at the same initial condition remain $O(\varepsilon)$-close on a discrete timescale of $O(1/\varepsilon)$. It

[*] Received by the editors July 11, 2003; accepted for publication (in revised form) by J. Meiss April 14, 2004; published electronically November 9, 2004. This work was supported by the DOE under grant DE-FG03-99ER41104.
http://www.siam.org/journals/siads/3-4/60043.html

[†] Department of Mathematics, University of Cincinnati, Cincinnati, OH 45221-0025 (Scott.Dumas@Math.uc.edu).

[‡] Department of Mathematics, University of New Mexico, Albuquerque, NM 87131 (ellison@math.unm.edu).

[§] Deutsches Elektronen Synchrotron, Notkestraße 85, 22607 Hamburg, Germany (vogtm@mail.desy.de).

is also often useful to use the continuous-time solutions of the corresponding averaged ODE

$$\frac{dy}{dt} = \varepsilon\widehat{f}(y) \tag{1.3}$$

to approximate the discrete-time solutions of (1.2) and hence also those of (1.1), so that we obtain the two approximation relations $x_n = y_n + O(\varepsilon)$ and $x_n = y(n) + O(\varepsilon)$ for $0 \leq n \leq O(1/\varepsilon)$ (note that y_n and $y(n)$ have different meanings). A more precise formulation appears below in Theorem 1, followed by a very elementary proof that makes no use of the usual transformation that appears in textbooks. (It is not always recognized that first-order averaging may be justified without a near-identity transformation of variables.)

Equation (1.1) is a special case of a more general problem on which we focus in this paper. Let $\nu \in \mathbf{R}$, $U \subseteq \mathbf{R}^d$, and $f : U \times \mathbf{R} \to \mathbf{R}^d$ be periodic with period 1 in its second argument. We then consider the system

$$x_{n+1} = x_n + \varepsilon f(x_n, n\nu). \tag{1.4}$$

The analysis of this problem is similar to the analysis of the flow problem $dx/dt = \varepsilon f(x,t)$ when f is quasiperiodic in t with two base frequencies, since small divisors enter both problems in the same way. Clearly (1.4) reduces to (1.1) when $\nu = q/p$ is rational. For ν irrational, we know from Weyl's equidistribution theorem that the average of $f(x, n\nu)$ over n exists and equals $\overline{f}(x) := \int_0^1 f(x,t)\,dt$. It is therefore natural to ask for what values of ν the solutions of (1.4) can be approximated by solutions of the two systems

$$y_{n+1} = y_n + \varepsilon\overline{f}(y_n) \tag{1.5}$$

and

$$\frac{dy}{dt} = \varepsilon\overline{f}(y). \tag{1.6}$$

In answering this question, it also seems natural (from the mathematical viewpoint) to introduce Diophantine conditions on ν, but these conditions in their usual form are problematic in applications, and not wholly necessary, as we shall see. In fact, we present approximation theorems that are both theoretically satisfying and suited to applications by weakening the usual small divisor conditions on ν (in which ν satisfies infinitely many Diophantine conditions), requiring instead only finitely many conditions at appropriately low order. These conditions exclude ν from zones centered on low-order rationals, and in this "far-from-low-order-resonance case" (where ν satisfies only "truncated Diophantine conditions" and is *not* necessarily irrational), we again find that $x_n = y_n + O(\varepsilon) = y(n) + O(\varepsilon)$ for $0 \leq n \leq O(1/\varepsilon)$ (see Theorem 2 below). Under the additional hypothesis that the average of the perturbation vanishes, we are able to show adiabatic invariance of solutions of system (1.4) on extended timescales up to $O(1/\varepsilon^2)$ (see Theorem 3). We thus have results both for low-order resonant (or rational) ν and for ν far from low-order resonance.

Finally, a simple procedure permits us to explore $O(\varepsilon)$ neighborhoods of low-order resonances $\nu = q/p$: we set $\nu = q/p + \varepsilon a$ (where the displacement parameter $a \in \mathbf{R}$) and rewrite

(1.4) as the system

$$\begin{pmatrix} x_{n+1} \\ \tau_{n+1} \end{pmatrix} = \begin{pmatrix} x_n + \varepsilon\, f(x_n, \frac{q}{p}n + \tau_n) \\ \tau_n + \varepsilon a \end{pmatrix}. \tag{1.7}$$

This is in the form of (1.1) with x_n replaced by $(x_n, \tau_n)^{\mathrm{T}}$. Writing $\widehat{f}(x, \tau) = 1/p \sum_{n=0}^{p-1} f(x, nq/p + \tau)$, the averaged problem reduces to

$$\begin{pmatrix} y_{n+1} \\ \tau_{n+1} \end{pmatrix} = \begin{pmatrix} y_n + \varepsilon\, \widehat{f}(y_n, \tau_n) \\ \tau_n + \varepsilon a \end{pmatrix}, \tag{1.8}$$

and we recapture the relations $x_n = y_n + O(\varepsilon) = y(n) + O(\varepsilon)$ for $0 \leq n \leq O(1/\varepsilon)$, where $y(t)$ is the solution of the system

$$\frac{d}{dt}\begin{pmatrix} y \\ \tau \end{pmatrix} = \varepsilon \begin{pmatrix} \widehat{f}(y, \tau) \\ a \end{pmatrix}, \tag{1.9}$$

which is equivalent to the nonautonomous system $dy/dt = \varepsilon \widehat{f}(y, \varepsilon a t)$; see Proposition C below. This method of exploring the neighborhoods of resonances generalizes to the case where ν (hence also a) is a vector (see [EDSV]); on the other hand, in higher dimensions it is also possible to use "resonant blocks" à la Nekhoroshev theory, as we do in [DE].

Initially, we state Theorems 1, 2, and 3 under the hypothesis that the perturbation εf has compact support in its x-domain, which is assumed to be all of $\mathbf{R}^d$; this avoids a priori restrictions on ε and permits clear proofs. To obtain results better suited to applications, we then give propositions that extend our theorems to more general perturbations on more general domains, and also to more general Diophantine conditions in which the zones mentioned above are allowed to depend on ε; this in turn allows ν to come within $O(\varepsilon^{\lambda})$ of low-order rationals, but with loss of accuracy in the approximation (see Propositions A and B below). Using the generalized versions of our theorems (provided by Propositions A, B, and C), we obtain an essentially complete description of solutions of system (1.4) on $O(1/\varepsilon)$ timescales for various values of ν. (However, there may be thin gaps at the boundaries between the ν for which the resonant and nonresonant normal forms (1.5), (1.6) and (1.8), (1.9) are valid; cf. Remark 2.7 below.)

From the viewpoint of applied mathematics, perhaps the most interesting aspect of our results is that Theorems 2 and 3 have physically realistic nonresonance conditions *in their hypotheses* yet provide approximations *valid on full $O(1/\varepsilon)$ time intervals.* In more general situations, such nice hypotheses lead to passage through resonance and thus to approximations that are valid only on somewhat shorter time intervals (cf. [ABG]); but we have identified an important class of simpler problems (arising, e.g., in accelerator beam dynamics) in which both the realistic hypotheses and the full $O(1/\varepsilon)$ validity times can coexist.

More generally, averaging principles for maps are not new; results in this direction have been available since the 1960's (cf., for example, [Bel],[Dr1],[Dr2]). However, a detailed theory of (1.4) suitable for applications appears to be missing from the literature, and we proceed in this paper to fill that gap. We do not, however, illustrate the full range of applicability of our theorems; instead we discuss an important class of problems which motivated this investigation, namely, the so-called kick-rotate models from accelerator dynamics, represented by $w_{n+1} = M\big(w_n + \varepsilon K(w_n)\big)$, where M is a stable linear symplectic map. The kick-rotate model takes the form of (1.4) under the transformation $w_n = M^n x_n$. We emphasize this model's application to the so-called weak-strong beam-beam (WSBB) interaction (see section 3.1 below), but kick-rotate models also apply to other localized perturbations in accelerators.

We stress that our discussion below in section 3 is the first mathematically rigorous treatment of this important class of models in the sense of asymptotics. There are of course many related discussions in the literature; beam dynamics treatments often begin with a smooth Hamiltonian formulation and apply canonical perturbation theory (see, e.g., [Ruth]), but without rigorous error analysis. Although perturbation theory is often viewed as inappropriate for treating resonances (cf. section 7 of [Ruth]), the paper [ES] examines both the nonresonant and resonant cases for flows using a perturbation theory complete with error bounds. However, the treatment of the small divisor problem there was crude; improvements along the lines of the present paper have since been developed and preliminary results are presented in [DE] and [DEV2]. Delta function perturbations are also often introduced into the (otherwise) smooth Hamiltonian framework, negating the possibility of error analysis and making the validity of approximations hard to assess. (The paper [CBW] gives a nice introduction to the beam-beam interaction and uses this Hamiltonian/delta function approach.) Formulating such problems as maps and rigorously analyzing the effect of delta function perturbations were important motivations for developing the methods in the present paper.

A notable exception to the smooth Hamiltonian treatment of beam dynamics is the work on maps using Lie operators, a good discussion of which may be found in [Fo], where the author carries this approach quite far—to realistic machine models—but without focusing on rigorous asymptotics. We are also aware of a research group working on highly mathematical perturbation treatments of beam dynamics in the context of maps [BGST], but our work here is quite distinct from theirs. Our perturbation parameter is the size of the "kick" (cf. section 3.1), whereas they study the long time stability of the origin (which is assumed to be a linearly stable elliptic fixed point), using the distance from the origin as a perturbation parameter. Futhermore, their analysis is quite complex, as they pursue Nekhoroshev-type results using many successive coordinate transformations which give rise to restrictive hypotheses that may be difficult to verify in practice. In our approach, resonances are treated in a simple yet rigorous way, and we obtain a natural partition of "tune space" into regions with distinct resonance properties. We believe this is an important new feature, both conceptually and practically. Of course, our method gives approximations to leading order only using no transformations; this accounts for much of its radical simplicity. It also allows us to use simple hypotheses and should eventually permit meaningful comparison of the kick-rotate approximation with numerical experiments. We have preliminary results on the size of the kicks (i.e., the values of ε) for which our approximations are valid, but we believe that this should be presented in the more realistic two-degree-of-freedom case (where ν is a vector). A progress report on this work appears in [EDSV]. We believe that our treatment provides the starting point for a simple effective means of studying mathematical models of beam dynamics rigorously and that its development will complement previous theoretical and mathematical work.

The remainder of this paper is organized as follows. In section 2 we present the details of our averaging results described informally above. In section 3 we apply the averaging principles to model problems in accelerator beam dynamics, showing that solutions of a class of "kick-rotate" models are well approximated by solutions of the corresponding averaged models. We also apply the adiabatic invariance principle to the Hénon map (often used to model sextupole magnets in accelerators). In section 4, we formulate the main technical tools required to prove the results in section 2. These are the so-called Besjes inequality for periodic functions (Lemma 1, section 4.1) and its generalization to functions far from low-order resonance (Lemma 2, section 4.2.2). After formulating and proving these inequalities, we use them to prove the mathematical results from section 2. Finally, in the appendix we give two elementary results used in earlier proofs.

We end this introduction with a few words about notation. We use the symbols $\mathbf{N}$, $\mathbf{Z}$, $\mathbf{R}_+$, and $\mathbf{R}$ to denote, respectively, the counting numbers $\{0, 1, 2, \dots\}$, the integers, the positive real numbers, and the real numbers. The symbol $|\ |$ indicates the Euclidean norm on $\mathbf{R}^d$ (or the absolute value $|k|$ of an integer k), and $\|\ \|_S$ denotes the uniform norm of a function over the set S; i.e., $\|F\|_S := \sup_{x\in S} |F(x)|$. We denote the open ball of radius $r > 0$ centered on x in $\mathbf{R}^d$ by $B_r(x)$. Given positive numbers T and ε, we define the $O(1/\varepsilon)$ discrete-time interval $N_{T/\varepsilon}$ by $N_{T/\varepsilon} := \{n \in \mathbf{N} \,\big|\, 0 \le n\varepsilon \le T\}$. Finally, in what follows we assume that all problems have been nondimensionalized so that $\varepsilon \ge 0$ is a dimensionless parameter; n and t are also dimensionless but are nevertheless referred to as "times."

2. Averaging principles and adiabatic invariance.

In this section we state and give brief remarks on our approximation results for maps.

2.1. Averaging for maps with periodic perturbations.

Let us be more precise about the functions f in (1.1) to which our results apply. Taking $S = U \times \mathbf{N}$, with $U = \mathbf{R}^d$ initially, we assume that $f : S \to \mathbf{R}$ satisfies the following:

(i) There exists a positive integer p such that $(x, n) \in S \Rightarrow f(x, n + p) = f(x, n)$.

(ii) f is bounded and locally x-Lipschitz on S.

(iii) There is an $r > 0$ such that $|x| \ge r$ and $n \in \mathbf{N} \Rightarrow f(x, n) = 0$.

When f satisfies (i), we say it is "periodic with period p in its second argument"; and when it satisfies (iii), it is "compactly supported in x, uniformly in n." By f locally x-Lipschitz on S, we mean that for every $x \in U$ there exist $\delta_x > 0$ and $L_x \ge 0$ such that $x_1, x_2 \in B_{\delta_x}(x)$ and $n \in \mathbf{N} \Rightarrow |f(x_2, n) - f(x_1, n)| \le L_x|x_2 - x_1|$. It follows from (ii) and (iii) that f is x-Lipschitz on S (i.e., there is an $L \ge 0$ such that $x_1, x_2 \in U$ and $n \in \mathbf{N} \Rightarrow |f(x_2, n) - f(x_1, n)| \le L|x_2 - x_1|$). In subsection 2.4 we show how to treat cases where U is an arbitrary open subset of $\mathbf{R}^d$ and f is not compactly supported.

We now state a simple averaging principle for maps with periodic perturbation $\varepsilon f(x, n)$.

Theorem 1. *Let $U = \mathbf{R}^d$, let $S = U \times \mathbf{N}$, suppose $f : S \to \mathbf{R}^d$ satisfies assumptions* (i), (ii), *and* (iii) *above with x-Lipschitz constant $L \ge 0$, and consider the system*

$$x_{n+1} = x_n + \varepsilon f(x_n, n) \tag{1.1}$$

together with the associated averaged systems

$$y_{n+1} = y_n + \varepsilon \widehat{f}(y_n) \tag{1.2}$$

and

$$\frac{dy}{dt} = \varepsilon \widehat{f}(y). \tag{1.3}$$

Then there is a nonnegative constant $C = C(f)$ such that the solutions x_n, y_n, and $y(t)$ of (1.1), (1.2), *and* (1.3) *with common initial condition $x_0 = y_0 = y(0) \in U$ exist uniquely for all time and satisfy $|x_n - y_n| \leq \varepsilon\, C\, p\, (1 + \varepsilon L)^n$ and $|x_n - y(n)| \leq \varepsilon\, C\, (1 + p)\, (1 + \varepsilon L)^n$.*

Remark 2.1. It is easy to see that the error bounds above are $O(\varepsilon)$ on $O(1/\varepsilon)$ time intervals as follows. Choose $T > 0$, and recall the definition (end of section 1 above) of the $O(1/\varepsilon)$ discrete time interval $N_{T/\varepsilon}$. Then for every $n \in N_{T/\varepsilon}$, $|x_n - y_n| \leq \varepsilon\, C\, p\, (1+\varepsilon L)^n \leq \varepsilon\, C\, p\, e^{\varepsilon L n} \leq \varepsilon\, C\, p\, e^{LT}$ and similarly $|x_n - y(n)| \leq \varepsilon\, C\, (1 + p) e^{LT}$. Clearly, these error bounds become transcendentally large on time intervals longer than $O(1/\varepsilon)$.

2.2. Averaging for maps with perturbations far from low-order resonance. We now present an averaging principle for system (1.4), where ν is a fixed positive number. When we write $\nu = q/p$, we mean that q and $p > 0$ are relatively prime integers with the *order* of the rational number ν given by p. Using this convention, we first note that if $\nu = q/p$, then $f(x, n\nu)$ has integer period p in n, and Theorem 1 applies. In fact, as shown by Proposition C, Theorem 1 applies not only *at* low-order rationals but also *near* them. But since the error bound in this theorem is proportional to p, it is not very useful when p is "large." We therefore restrict use of Theorem 1 to situations where p is "small" (the "low-order-resonance case"), and we next focus on situations where ν is far from low-order rational numbers (the "far-from-low-order-resonance case"). Small divisors inevitably enter the analysis (see the proof of Lemma 2, section 4.2.2) and it might be expected that ν would need to be "highly irrational" (e.g., satisfy infinitely many Diophantine conditions). We show instead that the averaging principle may be established when ν satisfies only finitely many Diophantine conditions to a certain order, and we call these *truncated Diophantine conditions.*

In more precise terms, ν satisfies truncated Diophantine conditions if it belongs to the set $\mathcal{D}(\phi, R)$ defined below in (4.3), where ϕ is the *zone function* of the Diophantine condition and $R \geq 1$ is the *truncation order* or *ultraviolet cutoff*, which gives precise meaning to the phrase "p large" used above (i.e., p is large if $p > R$). Roughly speaking, $\mathcal{D}(\phi, R)$ is constructed by removing open intervals centered on low-order rationals $\nu = q/p$. The zone function ϕ controls the size of the intervals removed, and the cutoff R is the maximal order of rationals from around which intervals are removed. These terms are defined precisely in subsection 4.2.1 (where we also discuss the difference between truncated and ordinary Diophantine conditions and indicate the advantages of the former; cf. section 4.2.1c).

We now introduce the class of functions to which our next result applies. Taking $S = U \times \mathbf{R}$ with $U = \mathbf{R}^d$ initially, we assume that $f : S \to \mathbf{R}^d$ satisfies the following (analogous to (i) through (iii) in section 2.1):

(j) $(x, \theta) \in S \Rightarrow f(x, \theta + 1) = f(x, \theta)$.

(jj) f is of class C^1 on S.

(jjj) There is an $r > 0$ such that $|x| \geq r$ and $\theta \in \mathbf{R} \Rightarrow f(x, \theta) = 0$.

Terminology for describing conditions (j) and (jjj) is similar to that for describing conditions (i) and (iii) above in subsection 2.1, and (jj) and (jjj) again imply that f is x-Lipschitz on S. Since we assume f has unit period in its second argument, its average $\overline{f}$ is simply $\overline{f}(y) := \int_0^1 f(y, \theta)\, d\theta$. Finally, we alert the reader that the zone function ϕ used below must be "adapted to f" in the sense that it must decay appropriately; this is made precise in (4.2) of subsection 4.2.1. (Basically ϕ must decay rapidly enough so that $\mathcal{D}(\phi, R)$ is nonempty but slowly enough so that the series in (4.2) converge. One way to ensure that such ϕ exists is to assume that $f \in C^4(S)$, as we show in part b of section 4.2.1. Nevertheless, we keep this requirement distinct from assumption (jj) above, since there may be other means of ensuring the existence of ϕ adapted to f; cf. Remark 4.3.)

We now state our averaging principle for maps (1.4) with ν far from low-order resonance as follows.

Theorem 2. *Let $U = \mathbf{R}^d$, let $S = U \times \mathbf{R}$, suppose $f : S \to \mathbf{R}^d$ satisfies assumptions* (j), (jj), *and* (jjj) *above with x-Lipschitz constant $L \geq 0$, and suppose the zone function ϕ is adapted to f on U in the sense of* (4.2). *Fix $\varepsilon \geq 0$, let R_ε be defined by* (4.7), *let $\nu \in \mathcal{D}(\phi, R_\varepsilon)$ as defined in* (4.3), *and consider the system*

$$x_{n+1} = x_n + \varepsilon f(x_n, n\nu) \tag{1.4}$$

together with the associated averaged systems

$$y_{n+1} = y_n + \varepsilon \overline{f}(y_n) \tag{1.5}$$

and

$$\frac{dy}{dt} = \varepsilon \overline{f}(y). \tag{1.6}$$

Then there exist nonnegative constants $C = C(f, \phi)$ and $C' = C'(f, \phi)$ such that the solutions x_n, y_n, and $y(t)$ of (1.4), (1.5), *and* (1.6) *with common initial condition $x_0 = y_0 = y(0) \in U$ exist uniquely for all time and satisfy $|x_n - y_n| \leq \varepsilon C (1 + \varepsilon L)^n$ and $|x_n - y(n)| \leq \varepsilon C' (1 + \varepsilon L)^n$.*

Remark 2.2. Remark 2.1 applies here as well. In addition, note that x_n depends on ν; however, y_n, $y(t)$, and the bounds are independent of ν for $\nu \in \mathcal{D}(\phi, R_\varepsilon)$.

Remark 2.3. It is also natural to consider the average $\lim_{N\to\infty}(1/N)\sum_{n=0}^{N-1} f(x, n\nu)$ of $f(x, n\nu)$ over n as mentioned in the introduction. Under mild integrability conditions on f, it can be shown that when ν is irrational, this average converges to $\int_0^1 f(x,\theta)\,d\theta$, which is the average used here (this is related to Weyl's equidistribution theorem; cf. [Br] and [Ko]). However, our results do not require the existence of the average of $f(x, n\nu)$ over n, nor do they require ν to be irrational; instead we require $\nu \in \mathcal{D}(\phi, R_\varepsilon)$, and this latter set contains (infinitely) many rationals of order greater than R_ε.

2.3. Adiabatic invariance on extended timescales.

In this subsection, we consider a special system somewhat like a perturbation of an integrable Hamiltonian system. As in Theorem 2, we assume that ν satisfies truncated Diophantine conditions, but now we assume additionally that the perturbation εf has zero mean; i.e., we assume that

(jw) for each $x \in \mathbf{R}^d$, $\int_0^1 f(x,\theta)\,d\theta = 0$.

This extra hypothesis gives an averaging principle showing that the (action-like) x variables are adiabatically invariant over timescales longer than $O(1/\varepsilon)$.

Theorem 3. *Let $U = \mathbf{R}^d$, let $S = U \times \mathbf{R}$, suppose $f : S \to \mathbf{R}^d$ satisfies assumptions* (j), (jj), (jjj), *and* (jw) *above, and suppose the zone function ϕ is adapted to f on U in the sense of*

(4.2). *Fix $\varepsilon \geq 0$, let R_ε be defined by* (4.7), *let $\nu \in \mathcal{D}(\phi, R_\varepsilon)$ as defined in* (4.3), *and consider the system*

$$x_{n+1} = x_n + \varepsilon f(x_n, n\nu) \tag{1.4}$$

with initial condition $x_0 \in U$. Then there exist nonnegative constants $K_1 = K_1(f, \phi)$ and $K_2 = K_2(f, \phi)$ such that the solution x_n of (1.4) *satisfies $|x_n - x_0| \leq K_1 \varepsilon + K_2 \varepsilon^2 n$ for $n \in \mathbf{N}$.*

Remark 2.4. We note that Theorem 3 shows that, given $T > 0$ and $0 \leq \alpha \leq 1$ and assuming $\varepsilon \leq 1$, we have $|x_n - x_0| \leq C(T)\, \varepsilon^\alpha$ for $0 \leq n \leq T/\varepsilon^{2-\alpha}$, where $C(T) = K_1 + K_2 T$. Using second (or higher) order averaging, it is possible to get a better estimate of $|x_n - x_0|$ on the full $O(1/\varepsilon^2)$ time interval (see [ES] for a flow version). Nevertheless, Theorem 3 is interesting because of its weak nonresonance conditions $\nu \in \mathcal{D}(\phi, R_\varepsilon)$ and because its proof is effected without the use of coordinate transformations (in fact, its proof is so short that it is nearly a corollary of Lemma 2 below; see section 4.2.4).

2.4. Extensions and generalizations. In this subsection we give three propositions that extend and generalize our results above, making them more suitable for applications. Our first proposition shows that Theorems 2 and 3 may be generalized to the case where the zones of the truncated Diophantine conditions depend on ε.

Proposition A (ε-dependent zone functions). *Suppose that $0 \leq \lambda \leq 1$ and that in Theorem 2 (or Theorem 3), the zone function ϕ is replaced by the new zone function $\varepsilon^\lambda \phi$, and R_ε is defined by* (4.7) *with ε replaced by $\varepsilon^{1-\lambda}$. Then the conclusions of the theorem remain true, provided that in the error bounds, $C\,\varepsilon$ and $C'\varepsilon$ are replaced by $C\varepsilon^{1-\lambda}$ and $C'\varepsilon^{1-\lambda}$ (or $K_1 \varepsilon + K_2 \varepsilon^2 n$ by $K_1 \varepsilon^{1-\lambda} + K_2 \varepsilon^{2-\lambda} n$ in Theorem 3).*

In order to clarify and simplify the mathematical structure of our methods, we have presented Theorems 1, 2, and 3 under the assumption that the perturbations have compact support on spatial domains U that are all of $\mathbf{R}^d$. Our next proposition shows that this assumption may be removed at little cost.

Proposition B (more general perturbations). *Suppose that in Theorems 1, 2, and 3 the first factor U of the domain S is an arbitrary open set $U \subset \mathbf{R}^d$ and that the compact support conditions* (iii) *and* (jjj) *are removed so that f is not necessarily (globally) x-Lipschitz with Lipschitz constant $L \geq 0$. Then the conclusions of Theorems 1 and 2 remain true provided that* (a) *the phrase "for all time" (governing existence, uniqueness, and approximation) is replaced by "$n \in N_{T/\varepsilon}$, i.e., for times not greater than T/ε," with $T > 0$ chosen strictly less than $\beta(x_0)$, where $[0, \beta(x_0))$ is the maximal forward interval of existence of the ε-independent averaged flow problem $dz/dt = \widehat{f}(z)$ (or $dz/dt = \overline{f}(z)$) in the domain U;* (b) $0 \leq \varepsilon < \varepsilon_0$*, where the threshold $\varepsilon_0(T) > 0$ is given below in section 4.3.2; and* (c) *the constant L appearing in the error bounds of Theorems 1 and 2 is replaced by L_D as defined in section 4.3.2. The conclusions of Theorem 3 remain true provided* (a) *the phrase "for all time" is replaced by "$n \in [0, T/\varepsilon^{2-\alpha}]$, i.e., for times not greater than $T/\varepsilon^{2-\alpha}$," where $T > 0$, $0 < \alpha \leq 1$ are parameters (as in Remark 2.4, except note $\alpha = 0$ is now excluded); and* (b) $0 \leq \varepsilon < \varepsilon_0$*, as above.*

Remark 2.5. Of course Proposition A also applies to Proposition B.

The following proposition shows that Theorem 1 may be used to analyze the dynamics of solutions of (1.4) in $O(\varepsilon)$ neighborhoods of low-order resonances $\nu = q/p$. An alternative to this approach appears in [DE].

Proposition C (behavior near low-order resonance). *Let $U \subset \mathbf{R}^d$ be open, let $S' = U \times \mathbf{R}$, and suppose $f : S' \to \mathbf{R}^d$ satisfies conditions* (j) *and* (jj) *of Theorem* 2 *with S replaced by S'. Fix the rational number q/p, $p > 0$ and q relatively prime, and fix $a \in \mathbf{R}$. Then* (1.4) *with $\nu = q/p + a\,\varepsilon$ may be rewritten as* (1.7)*, and Theorem* 1 *together with Proposition* B *apply with x and y replaced by $(x,\tau)^{\mathrm{T}}$ and $(y,\tau)^{\mathrm{T}}$, respectively. In particular there exist positive ε_0 and nonnegative L_D, $c = c(f,a)$, such that $|x_n - y_n| \le \varepsilon\, p\, c\,(1+\varepsilon L_D)^n$ and $|x_n - y(n)| \le \varepsilon\,(1+p)\,c\,(1+\varepsilon L_D)^n$ for $0 \le \varepsilon < \varepsilon_0$ and $n \in N_{T/\varepsilon}$.*

Remark 2.6. Clearly y_n evolves by $y_{n+1} = y_n + \varepsilon\widehat{f}(y_n, \varepsilon a n)$; and $y(n) = z(\varepsilon n)$, where $dz/dt = \widehat{f}(z, at)$.

Remark 2.7. Propositions A and B characterize the motion of x_n to within $O(\varepsilon^{1-\lambda})$ for ν far from low-order rationals, i.e., outside of $O(\varepsilon^{\lambda}\phi(p)/p)$ neighborhoods of rationals q/p with $0 < p \le R_\varepsilon$. For these ν the nonresonant normal form of (1.6) applies. Proposition C characterizes the motion to within $O(\varepsilon p)$ for ν inside $O(\varepsilon)$ neighborhoods of q/p. For these ν the resonant normal form of (1.9) applies. What may be missing is information about the motion for ν in gaps between the domains of validity of the normal forms of (1.5) and (1.7). The size of any gaps decreases to zero as $\lambda \nearrow 1$; but the error in the nonresonant normal form (1.5) simultaneously deteriorates to $O(1)$. High-order rationals, i.e., q/p with $p > R_\varepsilon$, are of course treated using Propositions A and B. It is interesting to note that they may also be treated using Proposition C; but the $O(\varepsilon p)$ error bound deteriorates to $O(1)$ as p approaches $O(1/\varepsilon)$.

3. Examples from accelerator beam dynamics.

Modern particle accelerators operate at the limits of current technology, and their design and operation depend crucially on understanding the dynamics of particle beams. In this section we give examples showing how Theorems 1 and 2 (supplemented by Propositions A, B, and C) may be used to analyze a class of beam dynamics models, and how Theorem 3 (similarly supplemented) may be used to analyze the Hénon map, which is itself a model of certain features in beam dynamics. Our averaging principles are especially effective for this purpose, as they compare solutions of the exact and averaged model problems in the simplest possible way and produce rigorous mathematical bounds on the nearness of these models' solutions. Although $O(1/\varepsilon)$ times are clearly the longest possible intervals on which nearness of individual trajectories is maintained in the general case, numerical simulations indicate that longer nearness times occur in the beam dynamics problems considered here. We furthermore expect action-like quantities to be adiabatically invariant on much longer $O(1/\varepsilon^2)$ timescales; formulating and proving such results mathematically is an important future goal and will almost certainly involve the results we present in this paper.

We point out that this section extends results of [ES] in at least two important ways: first, by using maps, we can incorporate delta function "kicks" that could not be treated rigorously via the flow methods of [ES]; second, the truncated Diophantine conditions used here are more physically realistic and explicit than the small divisor conditions of [ES] (cf. section 4.2.1 below). Finally, as in [ES], we do not require our maps to be polynomial here;

this is particularly important in applications to the WSBB problem.

3.1. The one-degree-of-freedom kick-rotate model. In this subsection, we focus on a simple but widely used class of beam dynamics models: the so-called one-degree-of-freedom "kick-rotate" models. As we proceed, we illustrate the important case of the weak-strong beam-beam (WSBB) interaction with explicit formulas. We also note that our methods may be generalized to treat models with several degrees of freedom and at higher order. (This will be the subject of a future publication [DESV]; a progress report appears in [EDSV].)

A circular accelerator (in storage mode) has a closed orbit; i.e., there is a unique solution of the equations of motion which has the periodicity of the accelerator. A complete three-degree-of-freedom description of single-particle dynamics involves three spatial coordinates in the comoving (Frenet–Serret) system, defined by the projection of the closed orbit on configuration space, and the three conjugate momenta. It is convenient to study the dynamics in terms of a Poincaré map (or one-turn map) at a fixed azimuthal location in the ring. Accelerators are designed to be as linear as possible, and thus transverse dynamics near the closed orbit can be modeled by a stable linear symplectic map with perturbations. Here we consider one transverse degree of freedom and define spatial and momentum coordinates, w_1 and w_2, so that the linear map is a rotation with "unperturbed tune ν of the so-called betatron motion." Perturbations of this model often consist of an instantaneous change in momentum w_2 depending only on the spatial coordinate w_1 at a fixed location in the ring (a "kick-map"). If we take this fixed location to be the azimuthal position of the Poincaré section, then the perturbed dynamics is given by the so-called kick-rotate model

$$w_{n+1} = R\left[w_n + \varepsilon\begin{pmatrix}0\\ -H'(w_{1,n})\end{pmatrix}\right],$$

$$\text{where}\quad R := e^{\mathcal{J}2\pi\nu}\quad\text{and}\quad \mathcal{J} := \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}, \tag{3.1}$$

i.e., a kick followed by rotation through angle $2\pi\nu$ about the origin. Here H' is the "kick function," and since R depends only on the fractional part of ν, we assume $\nu \in [0,1]$ in what follows. The map defined by (3.1) is symplectic since it is the composition of symplectic maps. The notation $w_{1,n}$ indicates the first component of the vector $w_n = (w_1, w_2)^{\mathrm{T}}_n$. (We hope the reader will forgive us the ambiguity of using w_n to denote a vector and w_1 or $w_{1,n}$ its first component and w_2 or $w_{2,n}$ its second component; the meaning should be clear from the context, since we rarely explicitly set $n = 1$ or $n = 2$.)

As a concrete example, we consider the WSBB effect for round Gaussian beams in collider rings; more details can be found in [DEV2]. The phase space distribution of the strong beam at the interaction point is assumed to be stationary; that is, the effect of the weak beam on the strong beam is ignored. Therefore, the beam-beam effect on the particle trajectories of the weak beam may be treated in the single particle picture, i.e., as a nonlinear kick due to the electromagnetic forces experienced while passing through a (longitudinally) short, time-independent, external charge distribution. We ignore coupling to the longitudinal motion, and we assume that the strong beam is represented by an axially symmetric charge distribution around the common closed orbit of the two beams in the transverse coordinate plane, so that it suffices to study a single phase plane. In this case, ε is a measure of the size of the beam-beam kick and $H'(w_1) := \big(1 - \exp(-w_1^2 r^{-2}/2)\big)/w_1$. By using the substitution $s^2/(2r^2) = w_1^2/(2r^2+s')$ one can show that

$$\begin{aligned} H(w_1) &:= \int_0^{w_1} \left(1 - \exp\left(-\frac{s^2}{2r^2}\right)\right)\frac{ds}{s} \\ &= \frac{1}{2}\int_0^{\infty} \left(1 - \exp\left(-\frac{w_1^2}{2r^2+s'}\right)\right)\frac{ds'}{2r^2+s'}, \end{aligned} \tag{3.2}$$

where we have taken $H(0) = 0$. Here r^2 is the ratio of the variances of the strong and weak beam Gaussians.

For $R = 1$, i.e., $\nu \in \{0,1\}$, (3.1) is easily solved and gives $w_n = (w_{1,0}, -nH'(w_{1,0}))^{\mathrm{T}}$ and thus $|w_{2,n}|$ is monotonically increasing to infinity. For $R = -1$ (i.e., $\nu = 1/2$), $w_{2n} = (w_{1,0}, -2nH'(w_{1,0}))^{\mathrm{T}}$ and the motion is again unbounded. Thus for $\nu \in \{0, 1/2, 1\}$ and for all initial conditions where $H'(w_{1,0}) \neq 0$, the distance from the origin is monotonically increasing. The basic question is, What happens for general ν? We shall apply the results of section 2 to answer this question for most ν in $[0,1]$.

Equation (3.1) may be written as $w_{n+1} = R\,w_n + \varepsilon\,R\,F(w_n)$, and the transformation $w_n = R^n\,x_n$ recasts this as

$$x_{n+1} = x_n + \varepsilon\,R^{-n}\,F(R^n\,x_n) =: x_n + \varepsilon\,f(x_n, n\nu), \tag{3.3}$$

which is in the standard form for averaging (cf. (1.4)).

It is easy to see that $f(x,\theta) = H'(x_1\cos 2\pi\theta + x_2\sin 2\pi\theta)\,(\sin 2\pi\theta, -\cos 2\pi\theta)^{\mathrm{T}} = (\partial H/\partial x_2, -\partial H/\partial x_1)^{\mathrm{T}}$. Thus if we define $\mathcal{H}(x,\theta) := H(x_1\cos 2\pi\theta + x_2\sin 2\pi\theta)$, then (3.3) becomes

$$x_{n+1} = x_n + \varepsilon\,\mathcal{J}\,\nabla_x\mathcal{H}(x_n, n\nu)\ . \tag{3.4}$$

Equations (3.3) and (3.4) also define symplectic maps, since the transformation is symplectic.

3.1.1. The kick-rotate model in the far-from-low-order-resonance case.

In this subsection, we examine the kick-rotate model (3.1) in the case where the tune belongs to the ε-dependent truncated Diophantine set $\mathcal{D}(\varepsilon^\lambda\phi, R_\varepsilon)$ (i.e., where the tune is "far from low-order resonance").

The most useful form of $\mathcal{H}$ in (3.4) is given in terms of the Fourier series $H(\sqrt{2J}\sin 2\pi t) = \sum_{k\in\mathbf{Z}} H_k(J)\,e^{i2\pi kt}$, from which it follows that $\mathcal{H}(x, n\nu) = \sum_{k\in\mathbf{Z}} H_k(J(x))\,e^{i2\pi k(\Phi(x)+n\nu)}$, where Φ and J are defined by $x_1 = \sqrt{2J}\sin(2\pi\Phi)$ and $x_2 = \sqrt{2J}\cos(2\pi\Phi)$. The averaged problem is then

$$y_{n+1} = y_n + \varepsilon\,\mathcal{J}\,\nabla_y H_0(J(y_n)), \tag{3.5}$$

where $H_0(J) = \int_0^1 H(\sqrt{2J}\sin 2\pi t)\,dt$. The ε-independent flow problem is

$$\frac{dz}{dt} = 2\pi\,\omega(J(z))\,\mathcal{J}z\ , \qquad z(0) = x_0, \tag{3.6}$$

where $2\pi\omega(J) = H_0'(J)$. We note that the map defined in (3.5) is only symplectic through $O(\varepsilon)$; however, the vector field in (3.6) has Hamiltonian $H(J(z))$. It is easy to check that $J(z) = \frac{1}{2}(z_1^2 + z_2^2)$ is constant along orbits so that $J(z) = J_0 = J(x_0)$ and thus $z(t) = e^{\mathcal{J}2\pi\omega(J_0)t}\,x_0$. Finally, Theorem 2 together with Propositions A and B give

$$w_n = e^{\mathcal{J}2\pi n(\nu+\varepsilon\omega(J_0))}\,x_0 + O(\varepsilon^{1-\lambda}) \tag{3.7}$$

for $n \in N_{T/\varepsilon}$, with ε suitably restricted to $0 \le \varepsilon < \varepsilon_0$ as in Proposition B for noncompactly supported perturbations, with $\lambda \in [0,1]$, $\nu \in \mathcal{D}(\varepsilon^\lambda\phi, R_\varepsilon)$, and with R_ε defined by the condition

$$\sum_{|k|>R_\varepsilon} \|H_k'(J)\nabla_x J\|_{D(\delta)} + \|H_k(J)2\pi\nabla_x\Phi\|_{D(\delta)} \;<\; \varepsilon, \tag{3.8}$$

where $D(\delta)$ is the δ-tube around the solution of (3.6) for $t \in [0,T]$. (The δ-tube is defined in section 4.3.2.)

From (3.7) the approximate motion is given once ω is known. In the WSBB case, $H_0(J) := \int_0^1 H(\sqrt{2J}\sin(2\pi t))\,dt = \frac{1}{2}\int_0^{J/(2r^2)}(1-e^{-w}I_0(w))\,\frac{dw}{w}$, where I_0 is the zeroth order modified Bessel function. (Note that $\exp(x\cos(y)) = I_0(x) + 2\sum_{k=1}^{\infty} I_k(x)\cos(ky)$.) Thus ω is given by

$$2\pi\omega(J) := H_0'(J) = \frac{1}{2J}\left(1 - \exp\left(-\frac{J}{2r^2}\right) I_0\left(\frac{J}{2r^2}\right)\right)$$

$$= \frac{1}{4\pi J}\int_0^{2\pi}\left(1 - \exp\left(-\frac{J\sin^2\vartheta}{r^2}\right)\right)d\vartheta. \tag{3.9}$$

The tune shift $\varepsilon\omega(J_0)$ is identical to that derived in [ES] and justifies the use of the delta function there.

3.1.2. The kick-rotate model in the near-to-low-order-resonance case. For ν near low-order resonance, we write $\nu = \frac{q}{p} + \varepsilon a$ when p is not too large (more precisely, when $0 < p \le R_\varepsilon$ for suitable $\varepsilon > 0$ in (3.8)). Thus using (1.7), our problem becomes

$$\begin{pmatrix} x_{n+1} \\ \tau_{n+1} \end{pmatrix} = \begin{pmatrix} x_n + \varepsilon\,\mathcal{J}\,\nabla_x\mathcal{H}(x_n, n\frac{q}{p} + \tau_n) \\ \tau_n + \varepsilon a \end{pmatrix}. \tag{3.10}$$

We are now in the periodic case, with averaged Hamiltonian $\widehat{\mathcal{H}}(x,\tau) = (1/p)\sum_{n=0}^{p-1} H(x_1\cos(2\pi \cdot[n\frac{q}{p}+\tau]) + x_2\sin(2\pi[n\frac{q}{p}+\tau]))$. The averaged problem is $(y_{n+1},\tau_n)^{\mathrm{T}} = (y_n + \varepsilon\,\mathcal{J}\,\nabla_y\widehat{\mathcal{H}}(y_n,\tau_n), \tau_n + \varepsilon a)^{\mathrm{T}}$, with its associated ε-independent flow $(dz/dt,\ d\tau/dt)^{\mathrm{T}} = (\mathcal{J}\nabla_z\widehat{\mathcal{H}}(z,\tau),\ a)^{\mathrm{T}}$. Solving for τ gives $\frac{dz}{dt} = \mathcal{J}\,\nabla_z\widehat{\mathcal{H}}(z,at)$. Theorem 1 with Propositions B and C then give $w_n = e^{\mathcal{J}2\pi n\nu}\,x_n = e^{\mathcal{J}2\pi n(\frac{q}{p}+\varepsilon a)}\,z(\varepsilon n) + O(\varepsilon)$ for $n \in N_{T/\varepsilon}$ and for $\nu = \frac{q}{p} + \varepsilon a$. However, it is not clear we have achieved a great simplification and so we look more closely. It turns out that $\widehat{\mathcal{H}}\,(\exp(-\mathcal{J}2\pi\theta')z, \theta) = \widehat{\mathcal{H}}(z, \theta - \theta')$, which suggests that an autonomous Hamiltonian system might be found with the symplectic transformation $z \mapsto \check{z}$ defined by $z = e^{-\mathcal{J}2\pi at}\,\check{z}$. This gives

$$\frac{d\check{z}}{dt} = 2\pi a\,\mathcal{J}\,\check{z} + \mathcal{J}\,\nabla_{\check{z}}\widehat{\mathcal{H}}(\check{z},0), \tag{3.11}$$

which indeed has autonomous Hamiltonian

$$\mathcal{K}(\check{z}) = 2\pi a J(\check{z}) + (1/p)\sum_{n=0}^{p-1} H(\check{z}_1\cos(2\pi nq/p) + \check{z}_2\sin(2\pi nq/p)). \tag{3.12}$$

The previous approximation thus becomes

$$w_n = e^{\mathcal{J}2\pi n\frac{q}{p}}\,\check{z}(\varepsilon n) + O(\varepsilon) \qquad \text{for} \qquad n \in N_{T/\varepsilon}, \tag{3.13}$$

from which the behavior of the approximation is now quite transparent.

In the WSBB case, $H(x)$ approaches a constant, $H(\infty)$, for large x. Thus $\mathcal{K}(\check{z})$ approaches $2\pi aJ(\check{z}) + H(\infty)$, and for $a \neq 0$ the integral curves become circles at large distances from the origin. The motion on these circles is clockwise for positive a and counterclockwise for negative a; thus a bifurcation in the phase plane portrait occurs at $a = 0$. In the case where $q/p \in \{0, 1/2, 1\}$ it is easy to see that $\widehat{\mathcal{H}}(\check{z}, 0) = H(\check{z}_1)$, and for $q/p \in \{1/4, 3/4\}$ one also easily finds $\widehat{\mathcal{H}}(\check{z}, 0) = 1/2\,[H(\check{z}_1) + H(\check{z}_2)]$ since H is an even function. For $q/p \in \{1/3, 2/3\}$ we find $\widehat{\mathcal{H}}(\check{z}, 0) = 1/3\,[H(\check{z}_1) + H(-\check{z}_1/2 + \sqrt{3}\check{z}_2/2) + H(-\check{z}_1/2 - \sqrt{3}\check{z}_2/2)]$. We briefly discuss the phase plane portraits for $\mathcal{K}$ in these cases (see [DEV1] and [DEV2] for figures).

In the first case ($q/p \in \{0, 1\}$) and for $a = 0$ we have $d\check{z}_1/dt = 0$ and $d\check{z}_2/dt = H'(\check{z}_{1,0})$. Thus the motion is identical to the exact case, as discussed just before (3.3), since (3.1) and (3.3) and the associated averaged problem are identical. For a small but positive, the origin is a (nonlinearly) stable center and the phase portrait is a one-parameter family of ovals which are long and thin in the $\check{z}_2$ and $\check{z}_1$ directions, respectively. As a increases to modest values the ovals become circular, consistent with the expectation of "stability far from low-order resonance." As a decreases from zero, the origin becomes a saddle, and two centers emerge from infinity at $(\pm c, 0)$, where $c \sim 1/\sqrt{2\pi|a|}$ for $|a|$ small. As a decreases further, the centers coalesce with the saddle at $4\pi ar^2 = -1$, and for $4\pi ar^2 < -1$ the only critical point is a center at the origin, again consistent with our expectation of stability.

The motion for $q/p = 1/2$ in the period two Poincaré map is identical to the motion for $q/p = 1$; the intermediate values may be obtained by rotating the phase plane portrait by a half turn.

For $q/p \in \{1/4, 3/4\}$ the phase plane portrait has four-fold symmetry, being invariant under reflections about the two axes and about the lines $\check{z}_2 = \pm\check{z}_1$. The origin is a critical point and its linearized vector field has eigenvalues $\pm 2\pi i(a - a_c)$, where $a_c = -1/(8\pi r^2)$. Thus the origin is a (nonlinearly) stable center for $a \neq a_c$, and it is easily checked that the origin is also a stable center for $a = a_c$ and that the rotation is clockwise for $a > a_c$ and counterclockwise for $a \leq a_c$. For $a \geq 0$ there are no other equilibria and the phase plane portrait is a one-parameter family of concentric ovals. For a small the (closed) integral curves look like four-pointed stars, with smoothed points on the axes, and as a increases the curves become circles. For $a_c < a < 0$ there are eight nonzero critical points. The four critical points $(\pm c, \pm c)$ are centers and the four at $(0, \pm c)$ and $(\pm c, 0)$ are saddle points, where c is the unique positive root of $4\pi ac + H'(c) = 0$. The critical points form an island structure in a neighborhood of radius c of the origin in the phase plane. This island structure emerges from infinity as a decreases through zero and coalesces in the origin as a decreases to a_c. For $a \leq a_c$, the origin is again the only equilibrium, and it is a stable center with counterclockwise rotation.

The portrait is again a one-parameter family of ovals approaching circles as a decreases from a_c.

Because H is an even function, $\widehat{\mathcal{H}}$ is the same for all $q/p \in \{1/6, 1/3, 2/3, 5/6\}$. Thus the phase plane portraits are the same for resonances of order three and six, and these portraits have a six-fold symmetry, being invariant under reflections about the axes $\check{z}_1 = 0$, $\check{z}_2 = 0$ and the lines $\check{z}_2 = \pm\check{z}_1/\sqrt{3}$, $\check{z}_2 = \pm\sqrt{3}\,\check{z}_1$. Qualitatively, the behavior as a function of a is similar to that in the case of resonance of order four (e.g., the island structure is similar, but there are now six rather than four islands). The critical value a_c at which the islands coalesce in the origin turns out to be the same as in the case $p = 4$.

3.1.3. Summary of the kick-rotate model.

We now have the following picture of the solutions of (3.1) on $O(1/\varepsilon)$ time intervals. For $\nu \in \mathcal{D}(\varepsilon^\lambda\phi, R_\varepsilon)$ the motion is given by (3.7), and thus our kick-rotate map behaves like a twist map with tune $\nu + \varepsilon\omega(J_0)$. For these ν the effect of the perturbation is slight; the up and down kicks on the integral curves essentially cancel, and the main effect of the perturbation is to create an amplitude-dependent tune. For $\nu = \frac{q}{p} + \varepsilon a$, we see that in the p-periodic Poincaré map, the approximate motion moves slowly along the phase curves given by the level curves of the Hamiltonian $\mathcal{K}(\check{z})$. As discussed for the WSBB case, this Hamiltonian has a rich variety of behaviors depending on the order p of the resonance and on the displacement $a\varepsilon$ from the resonance (in particular, the behavior varies considerably for $a > 0$, $a = 0$, and $a < 0$). We thus have an essentially complete picture of the motion (except for small gaps in ν as discussed in Remark 2.7).

3.2. The Hénon map.

We now apply Theorem 3 to the Hénon map (in beam dynamics this map is a standard model for the effect of a localized sextupole magnet in an otherwise linear lattice). The standard form of the Hénon map is (3.1) with $H(w_1) = w_1^3/3$. This gives (3.4) with $\mathcal{H}(x, \theta) = (x_1 \cos 2\pi\theta + x_2 \sin 2\pi\theta)^3/3$, which clearly has zero average. It follows that $f(x, \theta) = \mathcal{J}\,\nabla_x\mathcal{H}(x, \theta)$ in (1.4) has zero average, so that hypothesis (jw) of Theorem 3 is satisfied. Thus, by Theorem 3, Remark 2.4, and Proposition B, for appropriate $0 \le \varepsilon < \varepsilon_0$, $T > 0$, $\nu \in D(\phi, R_\varepsilon)$, and for any $0 < \alpha \le 1$, we have $|x_n - x_0| = O(\varepsilon^\alpha)$ on the discrete time interval $0 \le n \le T/\varepsilon^{2-\alpha}$.

Remark 3.1. The above discussion simply applies Theorem 3 as is (and thus also covers the case of more general H), but when $\mathcal{H}$ has a finite Fourier series (e.g., when H is a polynomial, as above) the proof of Theorem 3 may be simplified, both in terms of the smoothness requirement (see Remark 4.3) and in terms of the estimates in Lemma 2. In particular, for the Hénon map above, $g_k = 0$ except for $|k| \in \{1, 3\}$, so taking $R_\varepsilon = 3$, we see that the series defining C_1 and C_2 in Lemma 2 have only four terms each, while the tail-series of Lemma 2 vanishes.

4. Proofs and additional mathematical results.

As the title indicates, this is the most mathematical section of the paper. Subsection 4.1 treats periodic maps; this is quite straightforward and may be read as a kind of introduction to the deeper results of the next subsection. Subsection 4.2 concerns the considerably more complex case of maps far from low-order resonance and requires a (short) discussion of small divisors and truncated Diophantine conditions. The use of such conditions is not new, but as explained in the introduction, we believe our use of them in the present context is the most innovative aspect of this paper from the viewpoint of applied mathematics.

4.1. Periodic systems.

In this subsection we give a self-contained presentation of the remarkably simple technology required to prove the averaging principle for maps with periodic perturbations. This consists of the Besjes inequality for periodic functions (below), followed by its application to the proof of Theorem 1.

4.1.1. The Besjes inequality for periodic functions.

Let $U \subseteq \mathbf{R}^d$ be open, and let $S = U \times \mathbf{N}$. The Besjes inequality relies in an essential way upon the following assumption concerning the function $g : S \to \mathbf{R}$, periodic with period p in its second argument:

(iv) For each $y \in U$, $\sum_{n=0}^{p-1} g(y,n) = 0$.

When g has period p in n and satisfies (iv), we say it has *zero mean in n*. We now state the Besjes inequality for periodic maps as follows.

Lemma 1. *Let $U \subseteq \mathbf{R}^d$ be open and $S = U \times \mathbf{N}$ with $(y,n) \in S$. Suppose $g : S \to \mathbf{R}$ satisfies assumptions* (i) *(from section 2.1) and* (iv) *above and is bounded and y-Lipschitz on S with Lipschitz constant $L \geq 0$. If $\{y_n\}_{n=0}^{\infty} \subset U$ is a sequence for which the successive differences $y_{n+1} - y_n$ are bounded by M (i.e., $\sup_n |y_{n+1} - y_n| \leq M$), then for all $N \in \mathbf{N}$,*

$$\left|\sum_{n=0}^{N-1} g(y_n, n)\right| \leq \frac{1}{2} NpLM + p \|g\|_S.$$

Proof. Using the notation $[a]$ to designate the greatest integer in a, we first set $l = [(N-1)/p]$ (so that l is the number of periods of g contained in the segment $\{0, 1, 2, \dots, N-1\}$). Then using (iv), we write

$$\sum_{n=0}^{N-1} g(y_n, n) = \sum_{k=0}^{l-1} \sum_{n=0}^{p-1} \Big(g(y_{n+kp}, n) - g(y_{kp}, n) \Big) + \sum_{n=lp}^{N-1} g(y_n, n).$$

Now since g is y-Lipschitz and since $|y_{n+kp} - y_{kp}| \leq Mn$, we have

$$\begin{aligned}\left|\sum_{n=0}^{N-1} g(y_n, n)\right| &\leq \sum_{k=0}^{l-1} \sum_{n=0}^{p-1} LMn + \sum_{n=lp}^{N-1} |g(y_n, n)| \leq lLM \frac{p(p-1)}{2} + p \|g\|_S \\ &\leq \frac{1}{2} NpLM + p \|g\|_S. \qquad \blacksquare\end{aligned}$$

Remark 4.1. The original version of this lemma (Lemma 1 of [Bes]) was formulated for use in the proof of averaging principles for ODEs on $O(1/\varepsilon)$ timescales, and we use its analogue in a similar way below for maps. The original lemma assumes the time is at most $O(1/\varepsilon)$ and gives a final bound that is $O(\varepsilon)$, independent of time. We have found, however, that retaining the (here discrete) time-dependence makes the result more versatile (cf. the proof of Theorem 3 below).

We now illustrate the use of Lemma 1 by using it to prove Theorem 1.

4.1.2. Proof of Theorem 1.

Assume the hypotheses of Theorem 1 (cf. section 2.1). It is clear from assumption (iii) that the solutions x_n and y_n exist uniquely for all $n \in \mathbf{N}$. To see that the approximation relations hold, we write

$$|x_n - y_n| = \varepsilon \left| \sum_{k=0}^{n-1} \Big(f(x_k, k) - \widehat{f}(y_k) \Big) \right| \;=\; \varepsilon \left| \sum_{k=0}^{n-1} \Big(f(x_k, k) - f(y_k, k) + f(y_k, k) - \widehat{f}(y_k) \Big) \right|$$

$$\leq \varepsilon L \sum_{k=0}^{n-1} |x_k - y_k| + \varepsilon \left| \sum_{k=0}^{n-1} \widetilde{f}(y_k, k) \right|,$$

where $\widetilde{f}(y,n) := f(y,n) - \widehat{f}(y)$ is the "oscillating part of f." Let $g := \widetilde{f}$; then g satisfies the hypotheses of Lemma 1 with $U = \mathbf{R}^d$ and y-Lipschitz constant $2L$. From (1.2) and assumption (ii), we get $|y_{n+1} - y_n| \leq M := \varepsilon \|\widehat{f}\|_{\mathbf{R}^d}$, so Lemma 1 yields $\big|\sum_{k=0}^{n-1} \widetilde{f}(y_k, k)\big| \leq \frac{1}{2}\, n\, p\, 2L\varepsilon\, \|\widehat{f}\|_{\mathbf{R}^d} \;+\; p\, \|\widetilde{f}\|_S$, and thus $|x_n - y_n| \leq \varepsilon L \sum_{k=0}^{n-1} |x_k - y_k| + \varepsilon^2\, p\, L \|\widehat{f}\|_{\mathbf{R}^d} n + \varepsilon\, p \|\widetilde{f}\|_S$. Applying Lemma 3 (appendix) gives $|x_n - y_n| \;\leq\; \varepsilon p \left[(1 + \varepsilon L) \|\widehat{f}\|_{\mathbf{R}^d} + \|\widetilde{f}\|_S \right] (1 + \varepsilon L)^{n-1} \;\leq\; \varepsilon\, C\, p\, (1 + \varepsilon L)^n$ as claimed, where $C := \|\widehat{f}\|_{\mathbf{R}^d} + \|\widetilde{f}\|_S$. To prove the second inequality, we use Lemma 4 (appendix) to get $|y_n - y(n)| \leq \varepsilon \|\widehat{f}\|_{\mathbf{R}^d} (1 + \varepsilon L)^n$; then the triangle inequality gives $|x_n - y(n)| \leq \varepsilon \left[(1 + p) \|\widehat{f}\|_{\mathbf{R}^d} + p \|\widetilde{f}\|_S \right] (1 + \varepsilon L)^n \leq \varepsilon\, C\, [1 + p]\, (1 + \varepsilon L)^n$.

Remark 4.2. The preceding is no doubt one of the simplest possible proofs of an averaging principle for maps. Part of the simplicity derives from the use of Lemma 1, and part derives from the assumption of compact support (iii), which permits us to dispense with questions of the existence intervals for solutions. Thus, although assumption (iii) is often invalid in practice, by using it we are able to show that the basic estimates of the averaging method do not require restrictions on the size of ε; such restrictions are instead introduced by considering the solutions' existence intervals, or by methods of proof which rely on near-identity transformations (which may in turn require restrictions on ε for their inversion). Of course our results may be extended to cases with finite existence intervals (see Proposition B, section 2.4) and may also be combined with more traditional transformation methods to obtain results more efficiently at higher order [DESV].

4.2. Systems far from low-order resonance. In this subsection we generalize the Besjes inequality to functions far from low-order resonance and then use the generalization to prove Theorems 2 and 3. First, however, we present the following brief discussion.

4.2.1. Resonant zones, Diophantine conditions, and the ultraviolet cutoff. Here we discuss aspects of resonance, small divisors, and Diophantine conditions that will be needed in what follows. A more comprehensive introduction may be found in [Yo].

a. *Zone functions and Diophantine conditions.* In dynamical systems, Diophantine conditions arise naturally as a means of "controlling small divisors" and "avoiding resonances." In one dimension, divisors of the form $e^{2\pi i k \nu} - 1$ (with $0 \neq k \in \mathbf{Z}$ and $0 \neq \nu \in \mathbf{R}$) typically occur as denominators of terms in a series indexed over k, together with numerators which decrease to zero with increasing $|k|$. Clearly divisors cannot vanish, so rational (or "resonant") values of ν must be avoided. And although irrational ν do not cause divisors to vanish, when "nearly resonant," they may generate such small divisors as to cause divergence of the series in which they occur.

In order to be precise about avoiding small divisors, we introduce the concept of a *zone function* $\phi : \mathbf{R}_+ \to \mathbf{R}_+$, which is assumed to be *decreasing* (cf. the "approximation function" in [Ru1] and [Ru2]). We then define the "highly nonresonant" values of ν to be those belonging to the corresponding *Diophantine set*

$$\mathcal{D}(\phi) = \{\nu \in \mathbf{R} \,\big|\, |e^{2\pi i k\nu} - 1| \geq \phi(|k|),\ k \in \mathbf{Z}\backslash\{0\}\}, \tag{4.1}$$

which is a Cantor set. The Diophantine set $\mathcal{D}(\phi)$ may be thought of as $\mathbf{R}$ with countably many *zones* removed, where the zone $\mathcal{Z}_k = \{\nu \in \mathbf{R} \,\big|\, |e^{2\pi i k\nu} - 1| < \phi(|k|)\}$ corresponding to a particular $k \neq 0$ is a countable union of open intervals centered on rational numbers of the form q/k ($q \in \mathbf{Z}$). Further discussion of the structure of $\mathcal{D}(\phi)$ may be found in [BHS], or in [DEV2], where we indicate why a typical zone function of the form $\phi(r) = \gamma r^{-(\tau+1)}$ with $\gamma, \tau > 0$ removes zones of total length no more than $\gamma/(\pi\tau)$ from $[0, 1]$ (when this total length is less than one, the Diophantine set $\mathcal{D}(\phi)$ has positive measure and is therefore nonempty). Below we give conditions ensuring the existence of zone functions $\phi(r) = \gamma r^{-(\tau+1)}$ that work in our theorems. We are guided by the simple principle that ϕ must decrease at an appropriate rate: if ϕ decreases too slowly, then the union of the excluded zones may be so large that its complement, $\mathcal{D}(\phi)$, is empty; conversely, if ϕ decreases too rapidly, then $\mathcal{D}(\phi)$ may be too large and may contain values of ν so close to resonance as to cause divergence of the series in which small divisors appear.

The following terminology is useful for describing zone functions that decay appropriately. If $U \subseteq \mathbf{R}^d$ is open and $f : U \times \mathbf{R} \to \mathbf{R}^d$ has period 1 in its second argument and Fourier series $f(x, \theta) \sim \sum_{k\in\mathbf{Z}} f_k(x)\, e^{2\pi i k\theta}$ (where the kth Fourier coefficient is $f_k(x) = \int_0^1 f(x, \theta)\, e^{-2\pi i k\theta}\, d\theta$, requiring only that f is integrable in θ), then given a decreasing zone function ϕ, we say that ϕ is *adapted to f on $K \subseteq U$* provided

$$\mathcal{D}(\phi) \neq \emptyset, \qquad C_1(f, \phi) := \sum_{0\neq k\in\mathbf{Z}} \frac{\|f_k\|_K}{\phi(|k|)} < \infty, \qquad \text{and}$$

$$C_2(f, \phi) := \sum_{0\neq k\in\mathbf{Z}} \frac{\|Df_k\|_K}{\phi(|k|)} < \infty, \tag{4.2}$$

where $\mathcal{D}(\phi)$ is the Diophantine set of (4.1) and Df_k denotes the derivative of the function $f_k : U \to \mathbf{R}^d$ (and $\|Df_k\|_K$ denotes its induced uniform norm over K). Smoothness conditions on f ensuring the existence of zone functions adapted to f are not severe, as we now show.

b. *Smoothness conditions ensuring the existence of adapted zone functions.* Several questions naturally arise concerning the relationship between the smoothness of f and the existence of zone functions adapted to f as in (4.2). Formulating the sharpest possible conditions in this direction is somewhat delicate, but the following brief discussion should serve as a good starting point.

We first recall that for $\tau > 0$, the zone function $\phi(r) = \gamma r^{-(\tau+1)}$ generates a nonempty Diophantine set $\mathcal{D}(\phi)$ provided $\gamma > 0$ is sufficiently small (see [DEV2] or the more extensive discussion in section 1.2 of [BHS]). We assume that $f : U \times \mathbf{R} \to \mathbf{R}^d$ is of class $C^{p+1}(U \times \mathbf{R})$ and of compact support in the first argument, uniformly with respect to the second (cf. assumption (jjj) in section 2.2). Integrating the kth Fourier coefficient $f_k(x) = \int_0^1 f(x, \theta)\, e^{-2\pi i k\theta}\, d\theta$ by parts p times with respect to θ gives $f_k(x) = (2\pi i k)^{-p} \int_0^1 \big[\partial^p f/\partial\theta^p\big](x, \theta) e^{-2\pi i k\theta}\, d\theta$. Then taking the supremum over $x \in U$ of both sides of this expression gives $\|f_k\|_U \leq C(f,p)|k|^{-p}$, where $C(f,p) = \frac{1}{(2\pi)^p} \sup_{x\in U} \int_0^1 |\frac{\partial^p f}{\partial \theta^p}(x,\theta)|\, d\theta$. The same estimate holds for $\|Df_k\|_U$ with $C(f,p)$ replaced by $C'(f,p) = \frac{1}{(2\pi)^p} \sup_{x\in U} \int_0^1 |\frac{\partial^{p+1} f}{\partial x \partial \theta^p}(x,\theta)|\, d\theta$.

Using these estimates, we immediately deduce that both of the series in (4.2) are convergent provided that $p > \tau + 2$. Conversely, we see that whenever $p \geq 3$, there exists a zone function $\phi(r) = \gamma r^{-(\tau+1)}$ with $0 < \tau < p - 2$ which generates nonempty Diophantine sets $\mathcal{D}(\phi)$ (for γ sufficiently small) and which is adapted to f in the sense of (4.2). Thus one way to ensure the existence of zone functions adapted to f is to take f of class $C^4(U \times \mathbf{R})$.

Remark 4.3. A more refined (but lengthy) argument shows that the existence of ϕ adapted to f does not require quite as much smoothness as we demand above; we start our discussion under the assumption $f \in C^{p+1}(U \times \mathbf{R})$ primarily for simplicity. Of course, when f has a (sufficiently short) finite Fourier series, the decay rate of its terms is not an issue.

Remark 4.4. Although our results for system (1.4) as presented in this paper do not apply to the case of analytic perturbations εf (since analytic f with compact support vanishes identically), it would not be especially difficult to extend our theory to this case. For analytic $f : U \times \mathbf{T}^1 \to \mathbf{R}$ with Fourier coefficients f_k decreasing exponentially as, say, $\|f_k\|_U \leq \Gamma e^{-\beta|k|}$, it would be appropriate to use exponentially decreasing zone functions. In fact, given any $\rho > 0$, the zone function $\phi(r) = \gamma e^{-\rho r}$ generates nonempty Diophantine sets $\mathcal{D}(\phi)$ for small enough $\gamma > 0$. The decay rate β of the f_k must of course exceed ρ, which can be arranged provided f is analytic in its second argument with *analyticity parameter* $\alpha > \rho$. (This is an instance of the Paley–Wiener lemma; cf. [PW] or [BHS].) Roughly speaking, the analyticity parameter α is a measure of the minimum distance by which f may be extended as an analytic function on the complex torus (see also section 4.3.3 of [DEG] for an elementary discussion in the two-dimensional case).

It is interesting to note that Diophantine conditions corresponding to exponentially decaying ϕ may be strictly weaker than the weakest small-divisor conditions ordinarily used in dynamical systems, the so-called *Bruno conditions* (also spelled Brjuno or Bryuno; here "strictly weaker" means that the set $\mathcal{D}(\phi)$ properly contains the set of ν subject to Bruno conditions). This is, however, not surprising, since Bruno conditions apply to situations (such as conjugacies of circle diffeomorphisms, or KAM theory) in which countably many series with small divisors must simultaneously converge. By contrast, in Lemma 2 we require the convergence of only two series (in the language of [BHS], ours is a "one-bite" small-divisor problem).

c. *The ultraviolet cutoff and truncated diophantine conditions.* Finally, we introduce the notion of ultraviolet cutoff, which is important in physical applications of Diophantine conditions. To understand why, note that typically in applications, the ν that are required to be Diophantine are physical parameters. But checking whether a given ν belongs to a Cantor set of the form $\mathcal{D}(\phi)$ is a practical impossibility, since each point of $\mathcal{D}(\phi)$ has points arbitrarily close to it that are not in $\mathcal{D}(\phi)$. In other words, deciding if ν belongs to $\mathcal{D}(\phi)$ requires ν to be specified with infinite precision. Practically of course, it is only possible to specify physical parameters with finite precision. We surmount this difficulty by introducing *truncated Diophantine conditions* of the form

$$\mathcal{D}(\phi, R) = \{\nu \in \mathbf{R} \,\big|\, |e^{2\pi i k\nu} - 1| \geq \phi(|k|),\ k \in \mathbf{Z} \text{ with } 0 < |k| \leq R\}. \tag{4.3}$$

When $\nu \in \mathcal{D}(\phi, R)$, we say ν is *Diophantine to order* R *with respect to* ϕ, and we call R the *truncation order* or (*ultraviolet*) *cutoff*. $\mathcal{D}(\phi, R)$ is an approximating superset of $\mathcal{D}(\phi)$ with nonempty interior which converges to $\mathcal{D}(\phi)$ as $R \to \infty$. To decide if ν belongs to $\mathcal{D}(\phi, R)$, one checks only finitely many inequalities.

As a rough general rule, results in dynamical systems which are established for Diophantine sets $\mathcal{D}(\phi)$ may also be established (usually in slightly weaker form) for the corresponding larger, nicer sets $\mathcal{D}(\phi, R)$. The standard technique for doing so involves removing the "R-tail" of a series before applying Diophantine conditions and then checking that the tail is small. This technique was called the "ultraviolet cutoff" by Arnold in his proof of the KAM theorem and is illustrated in the proof of Lemma 2 below.

4.2.2. Besjes' inequality generalized to functions far from low-order resonance.

Lemma 2. *Let* $U \subseteq \mathbf{R}^d$ *be open,* $K \subseteq U$*,* $S = U \times \mathbf{R}$*, and suppose* $g : S \to \mathbf{R}^d$ *satisfies assumptions* (j), (jj) *from subsection* 2.2*, along with assumption* (jw) *from subsection* 2.3*. In addition, suppose* $\{y_n\}_{n=0}^{N_0-1} \subset K$ *is a sequence for which each line segment joining* y_n *to* y_{n+1} *is contained in* K*, and for which each successive difference* $y_{n+1} - y_n$ *is bounded by* M *(i.e.,* $\sup_n |y_{n+1} - y_n| \leq M$*); the zone function* ϕ *is adapted to* g *on* K *as in* (4.2)*; and the nonnegative constants* $C_1 = C_1(g, \phi)$ *and* $C_2 = C_2(g, \phi)$ *are defined by* (4.2)*. Let* $R \geq 1$ *and* $\nu \in \mathcal{D}(\phi, R)$*. Then for* $N < N_0$

$$\left|\sum_{n=0}^{N-1} g(y_n, n\nu)\right| \;\leq\; C_1 + N\left(C_2 M + \sum_{|k|>R} \|g_k\|_K\right),$$

$$\textit{where} \qquad \sum_{|k|>R} \|g_k\|_K \to 0 \quad \textit{as} \quad R \to \infty.$$

Proof. Since $C_1 < \infty$, we write g as its uniformly convergent Fourier series $g(y, \theta) = \sum_{0 \neq k \in \mathbf{Z}} g_k(x) e^{2\pi i k\theta}$, so that

$$\begin{aligned} \left|\sum_{n=0}^{N-1} g(y_n, n\nu)\right| &\leq \left|\sum_{n=0}^{N-1} \sum_{0<|k|\leq R} g_k(y_n) e^{2\pi i k n\nu}\right| \\ &\quad + \left|\sum_{n=0}^{N-1} \sum_{|k|>R} g_k(y_n)\, e^{2\pi i k n\nu}\right| \;=:\; I_N \,+\, II_N. \end{aligned} \tag{4.4}$$

We shall treat separately each of the double sums on the right-hand side of inequality (4.4). For the first double sum I_N, we reverse the order of summation and use the "summation by parts" formula $\sum_{n=0}^{N-1} a_n(b_{n+1} - b_n) = (a_N b_N - a_0 b_0) - \sum_{n=0}^{N-1} (a_{n+1} - a_n) b_{n+1}$ with $a_n = g_k(y_n)$

and $b_n = e^{2\pi ikn\nu}/(e^{2\pi ik\nu} - 1)$ so that $a_n(b_{n+1} - b_n) = g_k(y_n)e^{2\pi ikn\nu}$. It then follows that

$$\begin{aligned} I_N &\le \sum_{0<|k|\le R} \left| \frac{g_k(y_N)e^{2\pi iNk\nu} - g_k(y_0)}{e^{2\pi ik\nu} - 1} - \sum_{n=0}^{N-1} (g_k(y_{n+1}) - g_k(y_n)) \frac{e^{2\pi i(n+1)k\nu}}{e^{2\pi ik\nu} - 1} \right| \\ &\le \sum_{0<|k|\le R} \left(\frac{2\|g_k\|_K}{|e^{2\pi ik\nu} - 1|} + \frac{\|Dg_k\|_K}{|e^{2\pi ik\nu} - 1|} \sum_{n=0}^{N-1} |y_{n+1} - y_n| \right) \\ &\le \sum_{0<|k|\le R} \frac{2\|g_k\|_K + NM\|Dg_k\|_K}{|e^{2\pi ik\nu} - 1|} \\ &\le \sum_{0<|k|\le R} \frac{2\|g_k\|_K + NM\|Dg_k\|_K}{\phi(|k|)} \le \sum_{0\neq k} \frac{2\|g_k\|_K + NM\|Dg_k\|_K}{\phi(|k|)} = C_1 + NMC_2, \end{aligned} \tag{4.5}$$

where at the second inequality we have applied the mean value theorem to $g_k(y_{n+1}) - g_k(y_n)$ and thus $\|Dg_k\|_K$ is the induced norm. We treat the second double sum II_N of inequality (4.4) using the simple estimate

$$II_N \;\le\; N \sum_{|k|>R} \|g_k\|_K \;\to\; 0 \quad \text{as} \quad R \;\to\; \infty. \tag{4.6}$$

Inserting estimates (4.5) and (4.6) into inequality (4.4) concludes the proof. ■

Remark 4.5. A related analogous result for flows (but without the ultraviolet cutoff) appears as Lemma 13 of [Sa] and in Theorem 2 of [ES], and a more general Besjes-type inequality for so-called KBM vector fields also appears in [Sa] as Lemma 2. A still more closely related result for flows appears as Lemma 2 in [DEG], where it was used in averaging methods applied to certain charged particle motions in crystals.

Remark 4.6. In the case where g has a finite Fourier series, the above proof simplifies in obvious ways; but these simplifications become problematic as the Fourier series grows in length (note that the example in section 3.2 has a Fourier series with only four terms).

4.2.3. Proof of Theorem 2. Assume the hypotheses of Theorem 2 (cf. section 2.2). The proof is essentially the same as the proof of Theorem 1 with appropriate changes as needed in order to use Lemma 2. As in the previous proof, the solutions x_n and y_n clearly exist uniquely for all $n \in \mathbf{N}$. For the approximation relation, we write as in the proof of Theorem 1

$$|x_n - y_n| \le \; \varepsilon L \sum_{k=0}^{n-1} |x_k - y_k| \; + \; \varepsilon \left| \sum_{k=0}^{n-1} \widetilde{f}(y_k, k\nu) \right|.$$

The hypotheses clearly imply that $\|f\|_S < \infty$, and since ϕ is adapted to f on $K = U = \mathbf{R}^d$, the constants C_1 and C_2 from Lemma 2 (and (4.2)) are well defined. We take R_ε to be the smallest integer $R_\varepsilon \ge 1$ such that

$$\sum_{|k|>R_\varepsilon} \|f_k\|_K \le \varepsilon, \tag{4.7}$$

where $f_k(x)$ is the kth Fourier coefficient of f, and $K = U = \mathbf{R}^d$ (note that the inclusion $K \subset U$ is proper in certain other applications of inequality (4.7); cf. section 4.3.2). It is now a simple matter to check that if $\nu \in \mathcal{D}(\phi, R_\varepsilon)$, then the hypotheses of Lemma 2 are satisfied with $g = \widetilde{f}$, $K = U = \mathbf{R}^d$, $N_0 = \infty$, and $M = \varepsilon \|\overline{f}\|_{\mathbf{R}^d}$; thus $\big|\sum_{k=0}^{n-1} \widetilde{f}(y_k, k\nu)\big| \leq C_1 + n\big(C_2 \varepsilon \|\overline{f}\|_{\mathbf{R}^d} + \sum_{|k|>R_\varepsilon} \|f_k\|_{\mathbf{R}^d}\big) \leq C_1 + n\varepsilon\, C_2'$, where $C_2' := C_2 \|\overline{f}\|_{\mathbf{R}^d} + 1$. Applying Lemma 3 (in the appendix) gives $|x_n - y_n| \leq \varepsilon\big[C_1 + (\varepsilon + 1/L) C_2'\big]\,(1+\varepsilon L)^{n-1} \leq \varepsilon\,\big[C_1 + C_2'/L\big](1+\varepsilon L)^n =: \varepsilon\, C\,(1+\varepsilon L)^n$, as claimed. The second part of Theorem 2 again follows from Lemma 4 (appendix) and the triangle inequality. This gives $|x_n - y(n)| \leq \varepsilon\, C'(1+\varepsilon L)^n$, where $C' := C + \|\overline{f}\|_{\mathbf{R}^d}$.

Remark 4.7. It is important to note that for fixed positive ε, the ultraviolet cutoff R_ε need not be very large to ensure that inequality (4.7) holds, whence the number of inequalities to be checked in (4.3) (with $R = R_\varepsilon$) is also modest. In fact, straightforward estimation shows that when the Fourier coefficients of f decrease as $\|f_k\|_{\mathbf{R}^d} \leq C|k|^{-(p+1)}$ (e.g., when f is of class C^{p+1}), it is enough to take $R_\varepsilon \geq 1 + \big(\frac{2C}{p\varepsilon}\big)^{1/p}$. We also note that, in certain applications, it may be desirable to use a more refined version of (4.7) in which the right-hand side has an adjustable order constant (i.e., the ε on the right-hand side of (4.7) is replaced by $\zeta\varepsilon$, where ζ is a parameter; see [DEV2]).

4.2.4. Proof of Theorem 3.

The hypotheses of Theorem 3 (cf. section 2.3) ensure that the constants C_1, C_2 are well defined, so we set $K_1 = C_1$ and $K_2 = C_2 \|f\|_S + 1$. As in the proof of Theorem 2, we define R_ε by (4.7) and check as before that the hypotheses of Lemma 2 are satisfied (now $g = f$), from which we conclude that $|x_N - x_0| = \varepsilon\,|\sum_{n=0}^{N-1} f(x_n, n\nu)| \leq \varepsilon\, C_1 + \varepsilon N(C_2 M + \sum_{|k|>R_\varepsilon} \|f_k\|_{\mathbf{R}^d}) \leq \varepsilon\, C_1 + \varepsilon N(C_2 \varepsilon \|f\|_S + \varepsilon) = K_1 \varepsilon + K_2 \varepsilon^2 N$.

4.3. Proofs of Propositions A, B, and C.

(For statements of the propositions, see section 2.4.)

4.3.1. Proof of Proposition A.

The zone functions enter the proofs of Theorems 2 and 3 only through Lemma 2. It is clear that if ϕ is replaced by $\varepsilon^\lambda \phi$ in (4.5), then the estimate in (4.5) is changed to $\varepsilon^{-\lambda}(C_1 + NMC_2)$. If ε in (4.7) is replaced by $\varepsilon^{1-\lambda}$, then $\Big|\sum_{n=0}^{N-1} \widetilde{f}(y_n, n\nu)\Big| \leq \varepsilon^{-\lambda}(C_1 + N\varepsilon C_2')$ and the error bound in Theorem 2 then changes to $|x_N - y_N| \leq \varepsilon^{1-\lambda}(C_1 + C_2'/L)(1+\varepsilon L)^n =: \varepsilon^{1-\lambda} C(1+\varepsilon L)^n$, as claimed, while the error bound in Theorem 3 changes to $|x_N - x_0| \leq \varepsilon^{1-\lambda} C_1 + \varepsilon N\big(C_2 \varepsilon^{1-\lambda} \|f\|_S + \varepsilon\big) \leq K_1 \varepsilon^{1-\lambda} + K_2 \varepsilon^{2-\lambda} N$.

4.3.2. Proof of Proposition B.

The simplest case occurs in Theorem 3. Given $x_0 \in U$, choose positive $\delta < \operatorname{dist}(x_0, \partial U)$. Take $K = B_\delta(x_0) \subset U$ and note that ϕ is adapted to f on K by hypothesis. Let N_0 be the first exit time of x_n from K; thus Lemma 2 applies for positive $n < N_0$. Set $K_1 = C_1$ and $K_2 = 1 + C_2 \|f\|_{K \times \mathbf{R}}$, where C_1, C_2 are defined in (4.2), and choose $T > 0$ and $\alpha \in (0, 1]$. Then from the proof of Theorem 3, $|x_n - x_0| \leq K_1 \varepsilon + K_2 \varepsilon^2 n$ for $n < N_0$. If we require $0 \leq \varepsilon < \varepsilon_0(T) := \min\{\delta/2K_1, (\delta/2K_2 T)^{1/\alpha}\}$, then $x_n \in K$ and $|x_n - x_0| \leq K_1 \varepsilon + K_2 \varepsilon^2 n$ for $n \in [0, T/\varepsilon^{2-\alpha}]$.

We now give the proof as it applies to Theorem 2; the proof for Theorem 1 is similar (and simpler). As in the proof of Theorem 2, we consider the dynamics of (1.4), (1.5), and (1.6) with common initial condition $x_0 = y_0 = y(0) \in U$, but now we account for the solutions' possibly finite existence times by choosing the positive timescale parameter $T < \beta(x_0)$, where $[0, \beta(x_0))$ is the maximal forward interval of existence in the open set U of the ε-independent initial value problem

$$\frac{dz}{dt} = \overline{f}(z), \qquad z(0) = x_0 \in U. \tag{4.8}$$

Let $Z = \{z \in U \,|\, z = z(t),\ 0 \le t \le T\}$ denote the compact solution curve of system (4.8) over $[0, T]$ and choose positive $\delta < \operatorname{dist}(Z, \partial U)$. Then the compact closure $\overline{D}(\delta)$ of the open "δ-tube" $D(\delta)$ around Z formed by the union of open balls of radius δ having centers in Z is contained in U; i.e., $D(\delta) := \bigcup_{z\in Z} B_\delta(z) \subset \overline{D}(\delta) \subset U$. It follows that f is bounded and x-Lipschitz on $D(\delta) \times [0, 1]$ with x-Lipschitz constant $L_D \ge 0$ ($\overline{f}$ is similarly bounded and Lipschitz on $D(\delta)$, also with Lipschitz constant L_D). We note that $y(t) = z(\varepsilon t)$, and $(1 + \varepsilon L_D)^n \le e^{L_D T}$ for $n \in N_{T/\varepsilon}$. Using Lemma 4 (below; with $g = \overline{f}$, $V = D(\delta)$, and $L = L_D$), we see that the bound $|y_n - y(n)| \le \varepsilon(1 + \varepsilon L_D)^n \|\overline{f}\|_{D(\delta)}$ holds as long as y_n stays in $D(\delta)$. We thus require $\varepsilon < \delta/(2e^{L_D T}\|\overline{f}\|_{D(\delta)})$ so that $y_n \in B_{\delta/2}(y(n))$ for $n \in N_{T/\varepsilon}$. Since $|y_{n+1} - y_n| = \varepsilon|\overline{f}(y_n)| \le \varepsilon\|\overline{f}\|_{D(\delta)} < \delta/2$, it follows by the triangle inequality that both y_n and y_{n+1} are in $B_\delta(y(n))$. Since $B_\delta(y(n))$ is convex, the line segment joining y_n and y_{n+1} lies in $B_\delta(y(n))$, and hence in $D(\delta)$, so Lemma 2 applies to the sequence $\{y_n\}_{n\in N_{T/\varepsilon}}$ with $K = D(\delta)$. We then recover the estimates of Theorem 2 using this K and the L_D defined above. It follows that the bound $|x_n - y_n| \le \varepsilon\, C_3(1 + \varepsilon L_D)^n$ is valid as long as x_n stays in $D(\delta)$, where $C_3 = C_1 + (C_2\|\overline{f}\|_{D(\delta)} + 1)/L_D$, with C_1 and C_2 defined in (4.2). But $x_n \in D(\delta)$ for $n \in N_{T/\varepsilon}$ provided $\varepsilon < \delta/(2C_3 e^{L_D T})$. So we see finally that if $0 \le \varepsilon < \varepsilon_0 := \frac{1}{2}\delta e^{-L_D T} \min\{1/C_3,\ 1/\|\overline{f}\|_{D(\delta)}\}$, then $|x_n - y_n| \le \varepsilon\, C_3(1 + \varepsilon L_D)^n$ and $|x_n - y(n)| \le \varepsilon\big[\|\overline{f}\|_{D(\delta)} + C_3\big](1 + \varepsilon L_D)^n$ for $n \in N_{T/\varepsilon}$.

4.3.3. Proof of Proposition C. Let $g(u, n) := (f(x, nq/p + \tau),\ a)^{\mathrm{T}}$, where $u := (x, \tau)^{\mathrm{T}}$, so that $g : U \times \mathbf{R} \times \mathbf{N} \to \mathbf{R}^{d+1}$. The system $u_{n+1} = u_n + \varepsilon g(u_n, n)$ clearly satisfies the hypotheses of Proposition B applied to Theorem 1, with d replaced by $d + 1$ and U replaced by $U \times \mathbf{R}$. Thus Proposition B applied to Theorem 1 applies to u_n as well as to x_n. Futhermore,

$$\left|\begin{pmatrix} x_n \\ \tau_n \end{pmatrix} - \begin{pmatrix} y_n \\ \tau_n \end{pmatrix}\right| = |x_n - y_n| \qquad \text{and} \qquad \left|\begin{pmatrix} x_n \\ \tau_n \end{pmatrix} - \begin{pmatrix} y(n) \\ \tau(n) \end{pmatrix}\right| = |x_n - y(n)|;$$

thus $|x_n - y_n| \le \varepsilon\, p\, c\,(1 + \varepsilon L_D)^n$ and $|x_n - y(n)| \le \varepsilon\,(1 + p)\, c\,(1 + \varepsilon L_D)^n$. The constants L_D and $c = c(f, a)$ can be determined easily as in the proof of Theorem 1 supplemented by the proof of Proposition B.

Appendix. (Two elementary results with which the reader may be unfamiliar.)

Lemma 3 (Gronwall inequality for sequences). *Let A, α, and β be nonnegative and $\{E_n\}_{n=0}^{\infty}$ be a sequence of nonnegative real numbers satisfying $E_n \le A\sum_{k=0}^{n-1} E_k + \alpha n + \beta$. Then $E_n \le (AE_0 + \alpha + \beta + \alpha/A)(1 + A)^{n-1} - \alpha/A$.*

Proof. The proof is a short exercise in induction (the construction solves the inequality $R_n \le (1 + A)R_{n-1} + \alpha$, where $R_{n-1} = A\sum_{k=0}^{n-1}(E_k + \alpha n + \beta)$. ■

Lemma 4 (equivalence of autonomous flows and maps). *Let $V \subseteq \mathbf{R}^d$, $\varepsilon \ge 0$, and suppose $g : V \to \mathbf{R}^d$ is bounded and Lipschitz with Lipschitz constant $L \ge 0$. Then*

(1.5) *the map* $\qquad y_{n+1} = y_n + \varepsilon g(y_n)$

(1.6) *and the flow* $\qquad \dfrac{dy}{dt} = \varepsilon g(y)$

are equivalent in the sense that the solutions y_n and $y(t)$ of (1.5) *and* (1.6)*, respectively, with common initial condition $y_0 = y(0) \in V$ satisfy the nearness condition $|y_n - y(n)| \leq \varepsilon\|g\|_V(1+\varepsilon L)^n$ provided that y_k and $y(t)$ remain in V for $0 \leq k, t \leq n$.*

Proof. First we note that $y(n+1)-y(n) = \varepsilon\int_n^{n+1} g(y(t))\,dt = \varepsilon g(y(n)) + \varepsilon\int_n^{n+1}\big(g(y(t)) - g(y(n))\big)\,dt$. Thus $y_{n+1} - y(n+1) = y_n - y(n) + \varepsilon\big(g(y_n) - g(y(n))\big) - \varepsilon\int_n^{n+1}\big(g(y(t)) - g(y(n))\big)\,dt$. Setting $E_n = |y_n - y(n)|$, we obtain $E_{n+1} \leq E_n + \varepsilon L E_n + \varepsilon^2 L\|g\|_V$, since $\varepsilon|\int_n^{n+1}\big(g(y(t)) - g(y(n))\big)\,dt| \leq \varepsilon L\int_n^{n+1}|y(t) - y(n)|\,dt \leq \varepsilon^2 L\|g\|_V$. Using this last inequality to form a telescoping sum, we get $E_n - E_0 \leq \varepsilon L\sum_{k=0}^{n-1} E_k + n\varepsilon^2 L\|g\|_V$. Finally, we apply the Gronwall inequality above with $E_0 = 0$ to obtain $E_n \leq \varepsilon\|g\|_V(1+\varepsilon L)^n$. ■

Acknowledgments. The authors thank A. Bazzani, F. Golse, K. Heinemann, H. Mais, L. Michelotti, and T. Sen for stimulating discussions and the referees for valuable comments. H. Scott Dumas gratefully acknowledges support from the C.P. Taft Foundation at the University of Cincinnati.

REFERENCES

[ABG] M. Andreolli, D. Bambusi, and A. Giorgilli, *On a weakened form of the averaging principle in multifrequency systems*, Nonlinearity, 8 (1995), pp. 283–293.

[BGST] A. Bazzani, M. Giovannozzi, G. Servizi, G. Turchetti, and E. Todesco, *Resonant normal forms and stability analysis for area preserving maps*, Phys. D, 64 (1993), pp. 66–93.

[Bel] E.P. Belan, *On the averaging method in the theory of finite difference equations*, Ukraïn. Mat. Zh., 19 (1967), pp. 85–90 (in Russian).

[Bes] J. Besjes, *On the asymptotic methods for non-linear differential equations*, J. Mécanique, 8 (1969), pp. 357–372.

[BHS] H. W. Broer, G. B. Huitema, and M. B. Sevryuk, *Quasiperiodic Motions in Families of Dynamical Systems*, Lecture Notes in Math. 1645, Springer-Verlag, New York, 1996.

[Br] A. Browder, *Mathematical Analysis*, Springer-Verlag, New York, 1996.

[CBW] A. W. Chao, P. Bambade, and W. T. Weng, *Nonlinear beam-beam resonances*, in Nonlinear Dynamics Aspects of Particle Accelerators, Lecture Notes in Phys. 247, Springer-Verlag, New York, 1986, pp. 77–103.

[Dr1] V. A. Dragan, *Method of averaging systems of sum-difference equations*, Mat. Issled. 64 (1981), pp. 172–181, pp. 195–196 in Russian.

[Dr2] V. A. Dragan, *Method of averaging and freezing of systems of finite difference equations of two variables*, Mat. Issled., 64 (1981), pp. 182–188, pp. 196–197 (in Russian).

[DE] H. S. Dumas and J. A. Ellison, *Averaging for quasiperiodic systems*, in Proceedings of Equadiff 2003, J. Mawhin and S. Verduyn Lunel, eds., World Scientific, Singapore, to appear.

[DEG] H. S. Dumas, J. A. Ellison, and F. Golse, *A mathematical theory of planar particle channeling in crystals*, Phys. D, 146 (2000), pp. 341–366.

[DESV] H. S. Dumas, J. A. Ellison, T. Sen, and M. Vogt, *First Order Averaging Principle for Multifrequency Maps*, in preparation.

[DEV1] H. S. Dumas, J. A. Ellison, and M. Vogt, *First Order Averaging Principles for Maps with Applications to Beam Dynamics in Particle Accelerators*, presentation at 2002 Spring Meeting of APS in Albuquerque, NM, http://www.math.unm.edu/~ellison/papers/APS02_MAP.ps.gz.

[DEV2] H. S. Dumas, J. A. Ellison, and M. Vogt, *First-Order Averaging Principles for Maps with Applications to Beam Dynamics*, Preprint DESY 03-169, Hamburg, 2003, http://www-library.desy.de/report03.html.

[EDSV] J. A. Ellison, H. S. Dumas, M. Salas, T. Sen, A. Sobol, and M. Vogt, *Weak-strong beam-beam: Averaging and tune diagrams*, in Beam Halo Dynamics, Diagnostics, and Collimation: 29th ICFA Advanced Beam Dynamics Workshop on Halo Dynamics and Beam-Beam Interactions, AIP Conf. Proc. 693, J. Wei, W. Fischer, and P. Manning, eds., American Institute of Physics, New York, 2003, pp. 282–286.

[ES] J. A. Ellison and H.-J. Shi, *The method of averaging in beam dynamics*, in Accelerator Physics Lectures at the Superconducting Super Collider, AIP Conf. Proc. 326, Y. Yan and M. Syphers, eds., American Institute of Physics, New York, 1995, pp. 590–632.

[Fo] É. Forest, *Beam Dynamics: A New Attitude and Framework*, Harwood Academic Publishers, Amsterdam, 1998.

[Ko] T.W. Körner, *Fourier Analysis*, Cambridge University Press, Cambridge, UK, 1988.

[PW] R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, Amer. Math. Soc. Colloq. Publ. 19, AMS, Providence, RI, 1934.

[Ruth] R. D. Ruth, *Single particle dynamics and nonlinear resonances in circular accelerators*, in Nonlinear Dynamics Aspects of Particle Accelerators, Lecture Notes in Phys. 247, Springer-Verlag, Berlin, 1986, pp. 37–63.

[Ru1] H. Rüssmann, *On optimal estimates for the solutions of linear partial differential equations of first order with constant coefficients on the torus*, in Dynamical Systems, Theory and Applications, Lecture Notes in Phys. 38, Springer-Verlag, Berlin, 1975, pp. 598–624.

[Ru2] H. Rüssmann, *On the frequencies of quasi-periodic solutions of analytic nearly integrable Hamiltonian systems*, in Seminar on Dynamical Systems, Progr. Nonlinear Differential Equations Appl. 12, Birkhäuser, Basel, 1994, pp. 160–183.

[Sa] A. W. Sáenz, *Higher order averaging for nonperiodic systems*, J. Math. Phys., 41 (2000), pp. 5342–5368.

[Yo] J.-C. Yoccoz, *An introduction to small divisors problems*, in From Number Theory to Physics, Springer-Verlag, Berlin, 1992, pp. 659–679.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 4, pp. 433–462

Shear Dispersion along Circular Pipes Is Affected by Bends, but the Torsion of the Pipe Is Negligible[*]

A. J. Roberts[†]

Abstract. The flow of a viscous fluid along a curving pipe of fixed radius is driven by a pressure gradient. For a generally curving pipe it is the fluid flux which is constant along the pipe, and so I extend fluid flow solutions of Dean [W. R. Dean, *Phil. Mag.*, 5 (1928), pp. 673–695] and Topakoglu [H. C. Topakoglu *J. Math. Mech.*, 16 (1967), pp. 1321–1338] which assume constant pressure gradient. When the pipe is straight, the fluid adopts the parabolic velocity profile of Poiseuille flow; the spread of any contaminant along the pipe is then described by the shear dispersion model of Taylor [G. I. Taylor *Proc. Roy. Soc. London A*, 219 (1953), pp. 186–203] and its refinements. However, two conflicting effects occur in a generally curving pipe: viscosity skews the velocity profile which enhances the shear dispersion, whereas in faster flow centrifugal effects establish secondary flows that reduce the shear dispersion. The two opposing effects cancel at a Reynolds number of about 15. Interestingly, the torsion of the pipe seems to have very little effect upon the flow or the dispersion; the curvature is by far the dominant influence. Last, curvature and torsion in the fluid flow significantly enhance the upstream tails of concentration profiles in qualitative agreement with observations of dispersion in river flow.

Key words. shear dispersion, circular pipe, curvature, torsion

AMS subject classifications. 37N10, 37L10, 76M45, 76R50

DOI. 10.1137/030600886

1. Introduction.

Consider the dispersion of a contaminant material, with diffusivity D_c, in the steady laminar flow, velocity field $\boldsymbol{v}$, of a Newtonian fluid of density ρ and kinematic viscosity ν in an arbitrarily curving pipe of radius a such as the helical pipe shown in Figure 1. The material might be introduced as a localized release anywhere across the pipe or be fed in somehow at one end of the pipe. The flow is pumped by an overall pressure drop which maintains a fixed fluid flux along the pipe; that is, a constant mean velocity U is maintained irrespective of curvature. Nondimensionalize quantities with respect to the pipe radius a, the cross-pipe diffusion time a^2/D_c, and the reference pressure $\rho\nu U/a$. The Navier–Stokes and continuity equations for the steady, incompressible fluid flow then become

$$\mathrm{Re}\,\boldsymbol{v}\cdot\boldsymbol{\nabla}\boldsymbol{v} = -\boldsymbol{\nabla}p + \nabla^2\boldsymbol{v} \quad \text{and} \quad \boldsymbol{\nabla}\cdot\boldsymbol{v} = 0\,, \tag{1}$$

where $\mathrm{Re} = aU/\nu$ is the Reynolds number. The contaminant evolves according to the nondimensional advection-diffusion equation

$$\frac{\partial c}{\partial t} + \mathrm{Pe}\,\boldsymbol{v}\cdot\boldsymbol{\nabla}c = \nabla^2 c\,, \tag{2}$$

[*] Received by the editors September 10, 2003; accepted for publication (in revised form) by D. Lathrop May 24, 2004; published electronically November 9, 2004.
http://www.siam.org/journals/siads/3-4/60088.html

[†] Department of Mathematics & Computing, University of Southern Queensland, Toowoomba, Queensland 4350, Australia (aroberts@usq.edu.au).

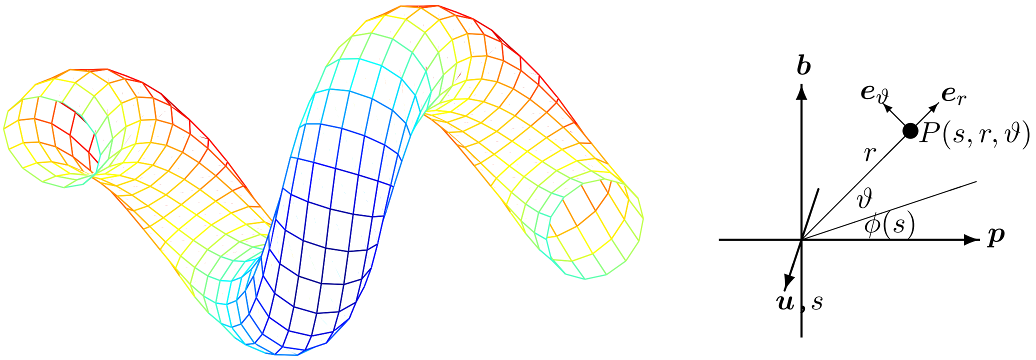


Figure 1. *Perspective drawing of a helical pipe as an example of the curving pipes containing fluid flow and contaminant dispersion that is modeled herein. On the right is a schematic diagram of the orthogonal curvilinear coordinate system local to the curving center line of the pipe: $\boldsymbol{u}$ points along the center line and s measures axial distance.*

where $\mathrm{Pe} = aU/D_c$ is the Peclet number. In typical liquids the Peclet number is much larger than the Reynolds number as their ratio, the Schmidt (or Prandtl) number $\mathrm{Sc} = \mathrm{Pe}/\mathrm{Re} = \nu/D_c$ is normally large (approximately 10^3 for the diffusion of material in liquids [33, p. 1119] although only about 8 for the diffusion of heat [1, p. 597][1]), whereas in typical gases the Schmidt number is roughly 1 and so the Peclet and Reynolds numbers are comparable. The analysis is interpreted with these two cases in mind of the Schmidt number being either $\mathcal{O}(1000)$ or $\mathcal{O}(1)$.

Here we analyze the flow and dispersion in an arbitrarily curving circular pipe. Most analysis of dispersion assumes a curved pipe is toroidal [33, 28, 19, 8], and most experiments are performed in helical pipes [38] (see further discussion by Berger, Talbot, and Yao [3]). Neglecting molecular diffusivity, dispersion in toroidal flow has been characterized from analytic formulae by Ruthven [33] and numerical solutions by McConalogue [25] using the residence time of different streamlines. Interestingly, using similar ideas, Jones and Young [20] deduced a regime of anomalous dispersion in a twisted but piecewise toroidal pipe. The fluid flow in helical pipes has been the subject of recent analysis [16, 39, 24, 42], whereas the flow in arbitrarily curving and twisting pipes has received little attention, although Pedley [29] accounted for leading order effects of variable curvature but ignored torsion, and Gammack and Hydon [14] investigated flow in pipes with exponentially varying curvature and torsion. Here I take these analyses further by simultaneously determining the fluid flow and the dispersion in arbitrarily curving circular pipes. The analysis is restricted to parameter regimes where the fluid flow is laminar—because of the induced secondary circulation, laminar flow is stable to higher Reynolds numbers in a curving pipe [38, 3]. The results thus will also be important in the flow and dispersion in microfluidic channels [17].

[1] The Prandtl number of water is 13.4 at 0°C, 9.5 at 10°C, 8.1 at 15°C, 7.1 at 20°C, 5.5 at 30°C, 4.3 at 40°C, 3.0 at 60°C, and 2.2 at 80°C, whereas the Prandtl number of air is 0.71 .

In section 2.1 we establish an orthogonal coordinate system based upon the arbitrarily curving geometry of the circular pipe, although requiring that the curvature of the center line varies smoothly. Let the center line of the pipe be described by $\boldsymbol{R}(s)$, where s measures arclength along the center line. Then a useful set of vectors for space in the vicinity of the pipe are the unit tangent $\boldsymbol{u}(s) = \boldsymbol{R}'$ of the center curve, unit normal $\boldsymbol{p}(s)$, and the unit binormal $\boldsymbol{b}(s)$ (see Figure 1). These vectors and the curvature $\kappa(s)$ and torsion $\tau(s)$ of the pipe are all connected by the Frenet formulae [22, section 8.7]

$$\boldsymbol{u}' = \kappa \boldsymbol{p}\,, \quad \boldsymbol{p}' = -\kappa \boldsymbol{u} + \tau \boldsymbol{b}\,, \quad \boldsymbol{b}' = -\tau \boldsymbol{p}\,. \tag{3}$$

Throughout this article I use a prime to denote $\partial/\partial s$. In a thin pipe the nondimensional velocity is approximately Poiseuille flow, $u \approx 2(1-r^2)$. However, there are corrections of $\mathcal{O}(\kappa)$ due to the curvature which are determined by solving the Navier–Stokes equations (1) for the fluid flow; see section 2, where low order expressions agree with the analysis of the flow in helical pipes by Tuttle [39].

Center manifold theory provides a rationale to form low-dimensional models of dynamics [32]. Here we model the long-time evolution of the large scale dispersion of contaminant along the pipe. Models of "long-waves" or "slowly varying in space" dynamics are justified in the center manifold approach, as outlined briefly in Appendix A, by requiring resolved longitudinal spatial structures to have small wavenumber or equivalently small gradients [30]. Within this slowly varying approximation, diffusion acts relatively quickly across the pipe to cause the contaminant concentration to be approximately constant in any cross-section: to leading order $c \approx C(s,t)$, where $C = \bar{c}$ is an average over a pipe cross-section (see Appendix A). The analysis, performed by the computer algebra of Appendix B, then systematically accounts for how variations along the pipe are affected by the varying velocity profile to disperse the contaminant. The center manifold theory (see Appendix A) asserts that, following an arbitrary release of material, this model of the dispersion along the pipe applies after transients on the time scale of cross-pipe diffusion. We thus report in section 3, as a generalization of Taylor's model [35, 36], that the advection-diffusion model

$$\frac{\partial C}{\partial t} \approx -\operatorname{Pe}\frac{\partial C}{\partial s} + \frac{\partial}{\partial s}\left(D\frac{\partial C}{\partial s}\right) \tag{4}$$

governs the large-scale dispersion of material along the pipe where for this case of a pipe of circular cross-section we find the effective diffusivity

$$\begin{aligned} D = {} & \left(1 + \frac{\kappa^2}{4}\right) \\ & + \frac{\operatorname{Pe}^2}{48}\left[1 + \kappa^2\left(\frac{863}{120} - \frac{7267\operatorname{Re}^2}{241920} + \frac{599\operatorname{Re}^4}{48384000} - \frac{2569\operatorname{Sc}^2\operatorname{Re}^4}{68428800}\right)\right] \\ & + \mathcal{O}(\kappa^4, \delta)\,. \end{aligned} \tag{5}$$

That is, we find the shear dispersion coefficient in a straight pipe, $D = \operatorname{Pe}^2/48$, is in a curved pipe modified by a factor approximately

$$1 + \kappa^2\left[7.2 - 3.0\left(\frac{\operatorname{Re}}{10}\right)^2 + (0.12 - 0.38\operatorname{Sc}^2)\left(\frac{\operatorname{Re}}{10}\right)^4\right].$$

As found by others, secondary circulation caused by fluid inertia depresses the effective dispersion by about $\kappa^2[3.0(\operatorname{Re}/10)^2+0.38\operatorname{Sc}^2(\operatorname{Re}/10)^4]$, but only for Reynolds numbers sufficiently large. In slow viscous flow, we find here that pipe curvature actually enhances the effective dispersion by about $7.2\kappa^2$—an effect that has apparently often been neglected [38, p. 317] in experimental determination of dispersion coefficients.

- For dispersion in gas flow, with $\mathcal{O}(1)$ Schmidt number Sc, the dispersion is depressed by secondary circulation only if the Reynolds number is greater than about 15, as otherwise the viscous enhancement is stronger. Remarkably, if the Schmidt number Sc is small, less than about 0.5, inertial effects in the fluid flow enhance the effective dispersion for Reynolds numbers greater than about 50.
- For dispersion in liquids, with, say $\mathcal{O}(1000)$ Schmidt number Sc, the term in $\kappa^2\operatorname{Re}^4\operatorname{Sc}^2$ dominates the other terms for the Reynolds number greater than about 0.5. Hence, in liquids and due to secondary circulations due to inertia, I reaffirm the reduction in effective dispersion.

Since the Dean number[2] $\operatorname{Dn}=2\sqrt{\kappa}\operatorname{Re}$, this last is the shear dispersion correction verified by Nunge, Lin, and Gill [28] and Johnson and Kamm [19] as being significant for $\operatorname{Dn}^2\operatorname{Sc}$ greater than about 100. A limitation of the expression (6) is that it predicts a physically unrealizable negative effective diffusivity for large enough Reynolds number. In most cases, Schmidt number Sc larger than 1, the term in $\kappa^2\operatorname{Sc}^2\operatorname{Re}^4$ dominates the correction. Hence to maintain a positive diffusion coefficient the Dean number $\operatorname{Dn}<25/\sqrt{\operatorname{Sc}}$ or equivalently $\operatorname{Sc}\operatorname{Dn}^2<650$. The expression (6) is a low order approximation to the correct curve, describing the downward curving shape on the left side of Figure 3 of Johnson and Kamm [19], but it needs the higher order corrections described in section 3.4 to describe the dispersion coefficient at a higher Dean number.

For a lower Reynolds number the qualitative deductions above vary from those of Nunge, Lin, and Gil [28] because their dispersion coefficient is different; see their equation (76). In particular, I predict that shear dispersion is frequently enhanced for gases, the reverse conclusion to that of Erdogan and Chatwin [11] and later Nunge, Lin, and Gil [28, pp. 363, 375]. I argue that the differences occur because all previous work, based upon the fluid flow solutions of Dean [10, 9] and Topakoglu [37], has assumed that the pressure *gradient* is fixed in the expansion in curvature κ—an adequate assumption for flow in a torus or helix where the curvature and the torsion are constant. But in a pipe of generally varying curvature and torsion, as developed here, it is the mean fluid flux which is fixed along the pipe and not the pressure gradient.[3] Since, for a constant pressure gradient the fluid flux varies with curvature and torsion—generally first decreasing with increasing torsion then later increasing with torsion (see Yamamoto, Yanase, and Yoshida [40] and its correction [41])—it follows that the mean pressure gradient (26) varies along a generally curving pipe. To check my

[2] As discussed by Berger, Talbot, and Yao [3, section 2.1.1.2], there are various and conflicting definitions of the Dean number: Berger, Talbot, and Yao recommended the use of $\operatorname{Dn}=2\sqrt{\kappa}\operatorname{Re}$, which I have adopted here. This Dean number could be viewed as $\sqrt{\kappa}\operatorname{Re}$ for a Reynolds number based upon the pipe diameter rather than the radius that I have used.

[3] Even Gammack and Hydon [14, p. 363] appear to fix the pressure gradient in their exponentially varying pipes by requiring the second order pressure correction $p_2\propto\sin\xi$, where ξ is their angular variable, and so their pressure correction has zero mean.

computer algebra program (listed in Appendix B) I temporarily fixed the pressure gradient in a helical pipe and found the resulting dispersion coefficient to be exactly equivalent to that given by Nunge, Lin, and Gill [28, equation (76)] except that the one term in $\mathrm{Re}^2\,\mathrm{Sc}\,\kappa^2$ (my κ is their $1/\lambda$) is zero in my results. I conjecture theirs is in error in this term. Because of the requirement to fix the fluid flux I recommend the use of (6) instead of the earlier published models of shear dispersion.

The error of $\mathcal{O}(\delta)$ in the shear dispersion coefficient given by (6) encompasses modifications due to torsion τ and to variations in curvature κ along the pipe: the parameter δ corresponds to the parameter η in Gammack and Hydon's analysis of exponentially varying pipes, $\kappa \propto e^{\eta s}$. The torsion affects only the dispersion coefficient at $\mathcal{O}(\kappa^2\tau^2)$, as seen in section 3, and so does not appear in (6). The effects of axial variations are reformulated as "memory" of the effective dispersion coefficient some distance upstream. Such memory effects in shear dispersion in varying channels were first recognized by Smith [34] and reflect the time it takes for the flow and the dispersion to relax to new conditions.

Using the computer algebra of Appendix B I straightforwardly determine both higher order corrections to the dispersion coefficient (section 3.4) and high order terms in the advection-diffusion equation itself (section 3.3). These terms may be used either to refine the approximations or to give good estimates of the errors in a lower order approximation. Earlier work by Mercer and Roberts [26] gave a sharp estimate for the limit of spatial resolution in a straight circular pipe.

2. The fluid flow.

The first task is to find the laminar viscous fluid flow in the curving pipe. The focus of the paper is the dispersion in the pipe by the flow, but there are enough interesting and relevant features in the fluid flow itself to be discussed briefly here—in particular, this section confirms aspects of my analysis by reproducing many results of other authors about steady laminar flow in curved and twisted pipes.

We assume that the flow is steady as appropriate to flow driven by a constant pressure drop through a fixed pipe. However, to maintain everywhere constant fluid flux, the mean pressure gradient, $\bar{p}'$, varies with the curvature of the pipe as given in (26).

2.1. The orthogonal curvilinear coordinate system.

Expressions for the flow are derived in an orthogonal curvilinear coordinate system matched to the geometry of the circular pipe. The orthogonal coordinate system has been used by Germano [16], Kao [21], Liu and Masliyah [24], and Yamamoto, Yanase, and Yoshida [40, 41] to investigate the structure of the fluid flow in helical pipes up to Dean numbers of 2000 and is well known in hydromagnetodynamics. One difference here is that we do not assume the pipe is helical; instead we allow arbitrary variations in the curvature and torsion of the pipe—the one important restriction is that the curvature and torsion must vary only slowly along the tube. Such slow variations along the pipe were also assumed by Murata, Liyake, and Inaba [27] in their analysis of the flow in tubes bent sinusoidally in a plane and Pedley [29] in a leading approximation to the effects of curvature. As shown schematically in Figure 1, positions in space are labeled by (s, r, ϑ) and have position vector

$$\boldsymbol{r} = \boldsymbol{R}(s) + r\cos\theta\,\boldsymbol{p} + r\sin\theta\,\boldsymbol{b}\,, \quad \text{where} \quad \theta = \vartheta + \phi(s) \tag{6}$$

measures the angle from the plane of the normal $\boldsymbol{p}$ to the point $\boldsymbol{r}$; thus $\theta = 0$ corresponds to the inside of the local bend, whereas $\theta = \pm\pi$ corresponds to the outside; and where the nondimensional radius ranges over $0 \le r \le 1$. However, due to torsion in the shape of the pipe the reference plane of the orthogonal coordinate system must twist along the pipe by an amount $\phi(s)$, where

$$\frac{d\phi}{ds} = -\tau\,. \tag{7}$$

The unit vectors and scale factors of this orthogonal coordinate system are then

$$\begin{aligned} h_s &= 1 - \kappa r\cos\theta\,, & \boldsymbol{e}_s &= \boldsymbol{u}\,,\\ h_r &= 1\,, & \boldsymbol{e}_r &= \cos\theta\,\boldsymbol{p} + \sin\theta\,\boldsymbol{b}\,,\\ h_\vartheta &= r\,, & \boldsymbol{e}_\vartheta &= -\sin\theta\,\boldsymbol{p} + \cos\theta\,\boldsymbol{b}\,. \end{aligned} \tag{8}$$

Note that all expressions for fluid and concentration fields are written in terms of θ, the angle relative to the local direction of curvature of the bent pipe, because it is this angle that primarily determines the shape of the local fields, but all equations are written in terms of the angular coordinate in the orthogonal system, namely, ϑ; remember that θ varies with ϑ and s according to (6). Observe the scale factors are all positive provided $0 < r < 1/\kappa$, and so the coordinate system is well defined for unit radius pipes provided the nondimensional center line curvature $\kappa < 1$. Let the velocity field, with components the axial velocity u, the radial velocity v, and the angular velocity w, be denoted by

$$\boldsymbol{v} = u\boldsymbol{e}_s + v\boldsymbol{e}_r + w\boldsymbol{e}_\vartheta\,.$$

Then, noting that it is convenient to compute the viscous dissipation term via the vorticity (as does Tuttle [39, p. 548]),

$$\nabla^2\boldsymbol{v} = -\boldsymbol{\nabla}\times\boldsymbol{\omega}\,,\quad \boldsymbol{\omega} = \boldsymbol{\nabla}\times\boldsymbol{v}\,,$$

standard formulae apply for computing components of the Navier–Stokes equations (1) [1, Appendix B]:

$$\omega_s = \frac{1}{r}\left(\frac{\partial(rw)}{\partial r} - \frac{\partial v}{\partial\vartheta}\right), \tag{9}$$

$$\omega_r = \frac{1}{rh_s}\left(\frac{\partial(h_s u)}{\partial\vartheta} - \frac{\partial(rw)}{\partial s}\right), \tag{10}$$

$$\omega_\vartheta = \frac{1}{h_s}\left(\frac{\partial v}{\partial s} - \frac{\partial(h_s u)}{\partial r}\right), \tag{11}$$

$$\begin{aligned} 0 = {}& \frac{1}{h_s}\frac{\partial p}{\partial s} + \frac{1}{r}\left(\frac{\partial(r\omega_\vartheta)}{\partial r} - \frac{\partial\omega_r}{\partial\vartheta}\right)\\ &+ \operatorname{Re}\left(\frac{u}{h_s}\frac{\partial u}{\partial s} + v\frac{\partial u}{\partial r} + \frac{w}{r}\frac{\partial u}{\partial\vartheta} + \frac{uv}{h_s}\frac{\partial h_s}{\partial r} + \frac{uw}{rh_s}\frac{\partial h_s}{\partial\vartheta}\right), \end{aligned} \tag{12}$$

$$\begin{aligned} 0 = {}& \frac{\partial p}{\partial r} + \frac{1}{rh_s}\left(\frac{\partial(h_s\omega_s)}{\partial\vartheta} - \frac{\partial(r\omega_\vartheta)}{\partial s}\right)\\ &+ \operatorname{Re}\left(\frac{u}{h_s}\frac{\partial v}{\partial s} + v\frac{\partial v}{\partial r} + \frac{w}{r}\frac{\partial v}{\partial\vartheta} - \frac{w^2}{r} - \frac{u^2}{h_s}\frac{\partial h_s}{\partial r}\right), \end{aligned} \tag{13}$$

$$0 = \frac{1}{r}\frac{\partial p}{\partial \vartheta} + \frac{1}{h_s}\left(\frac{\partial \omega_r}{\partial s} - \frac{\partial (h_s\omega_s)}{\partial r}\right) + \operatorname{Re}\left(\frac{u}{h_s}\frac{\partial w}{\partial s} + v\frac{\partial w}{\partial r} + \frac{w}{r}\frac{\partial w}{\partial \vartheta} - \frac{u^2}{rh_s}\frac{\partial h_s}{\partial \vartheta} + \frac{vw}{r}\right), \tag{14}$$

$$0 = \frac{1}{rh_s}\left(\frac{\partial (ru)}{\partial s} + \frac{\partial (rh_s v)}{\partial r} + \frac{\partial (h_s w)}{\partial \vartheta}\right). \tag{15}$$

See the curvature of the pipe enters these fluid flow equations predominantly through the variations of the scale factor h_s (see (8)) both across and along the pipe. The torsion of the pipe appears more subtly through the difference between θ, the angle relative to the local direction of curvature, and ϑ, the angle in the coordinate system; thus spatial derivatives of θ nontrivially depend upon the torsion through (6)–(7). The above steady Navier–Stokes equations are solved with a fixed fluid flux and with zero velocity on the pipe walls: $u = v = w = 0$ on $r = 1$. The computer algebra program in Appendix B solves these equations iteratively.

There are some subtleties in the geometry of the coordinate system. As discussed by Zabielski and Mestel [42, section 2.2], observe that because of the twist in a helical pipe the axial unit vector $\boldsymbol{u}$ is *not everywhere* tangent to the lines of helical symmetry—the s-coordinate curves are not curves of helical symmetry. Thus be careful in interpreting cross-flow velocities v and w because in one view they will involve a small component of the relatively large velocity along the lines of helical symmetry. In an alternative presented by Tuttle [39], the twist in the coordinate system caused by torsion generates an effect similar to that caused by a coordinate system rotating in time. However, here we consider flow in a generally curving pipe with no large scale symmetry, so the only definite longitudinal direction is the local unit vector $\boldsymbol{u}$, and we thus discuss v and w as cross-flow velocities, as do Gammack and Hydon [14]. Similarly, in helical symmetry one cannot find a cross-section *plane* normal to the lines of helical symmetry [42, p. 300] so an arbitrary decision is needed. As is conventional for helical pipes and as simplest for generally curving pipes, we conventionally take a cross-section to be normal to the centerline of the pipe. In these cross-sections the pipe is circular.

2.2. Stokes flow. Solving the fluid equations using the computer algebra program in Appendix B, I deduce the Stokes flow field, Re = 0, is

$$u = (1-r^2)\left[2 + \kappa\tfrac{3}{2}r\cos\theta + \kappa^2\tfrac{5}{8}r^2\cos 2\theta - \kappa^2\tfrac{11}{48}(1-3r^2)\right] + \mathcal{O}(\kappa^3, \delta^2, \operatorname{Re}), \tag{16}$$

$$v = \tfrac{1}{3}(1-r^2)^2\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] + \tfrac{1}{96}r(1-r^2)^2\left[(38 + 43\cos 2\theta)\kappa\kappa' + 43\sin 2\theta\,\kappa^2\tau\right] + \mathcal{O}(\kappa^3, \delta^2, \operatorname{Re}), \tag{17}$$

$$w = \tfrac{1}{6}(1-r^2)(2-r^2)\left[-\sin\theta\,\kappa' + \cos\theta\,\kappa\tau\right] + \tfrac{1}{96}r(1-r^2)\left[(43-29r^2)(\cos 2\theta\,\kappa^2\tau - \sin 2\theta\,\kappa\kappa') + (6-2r^2)\kappa^2\tau\right] + \mathcal{O}(\kappa^3, \delta^2, \operatorname{Re}), \tag{18}$$

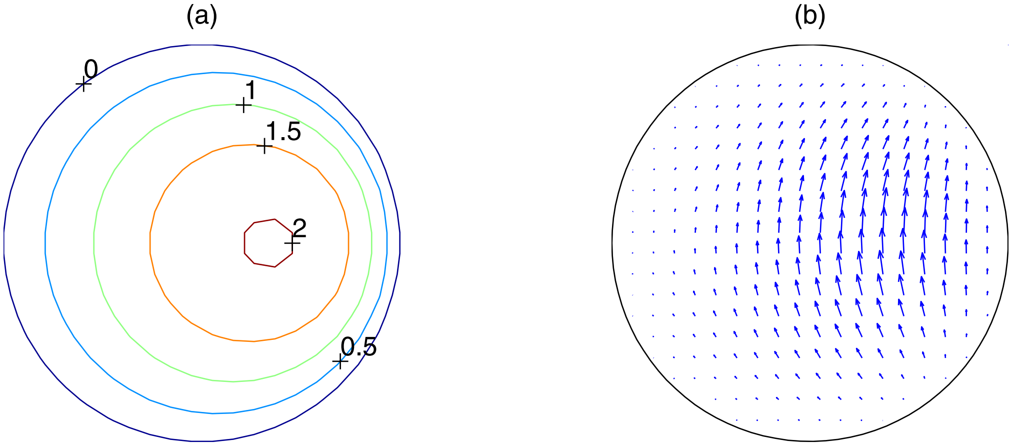


Figure 2. (a) *Contours of axial velocity u of the viscously dominated Stokes flow in a helical pipe with curvature $\kappa = 0.8$ (chosen so large to accentuate the modifications).* (b) *Corresponding torsion induced cross-pipe fluid velocities to leading order in the torsion. The plots are evaluated from the asymptotic solution with errors $\mathcal{O}(\kappa^5)$.*

$$\bar{p}' = -8 + \tfrac{1}{6}\kappa^2 + \mathcal{O}(\kappa^3, \delta^2, \mathrm{Re}) \,, \tag{19}$$

$$\begin{aligned} p = \bar{p} &- \tfrac{1}{3}r(1-3r^2)\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] \\ &- \tfrac{1}{24}(5+4r^2-21r^4)\kappa\kappa' - \tfrac{1}{24}r^2(9-26r^2)\left[\cos 2\theta\,\kappa\kappa' + \sin 2\theta\,\kappa^2\tau\right] \\ &+ \mathcal{O}(\kappa^3, \delta^2, \mathrm{Re}) \,, \end{aligned} \tag{20}$$

where δ is used to denote the order of magnitude of derivatives of the quantities varying slowly along the pipe. For example, κ' and $\tau = -\phi'$ are thus $\mathcal{O}(\delta)$.

See that, for example, the Stokes flow in a torus ($\kappa = \text{const}$ and $\tau = 0$) is simply one of axial flow (see Figure 2(a)) in an adjusted mean pressure gradient as all other components vanish. The axial velocity maximum is shifted to the inside of the curve (to the right in Figure 2(a)) and is increased slightly. In contrast to flows at a significant Reynolds number, the pressure gradient around a curve is less than that in a straight pipe presumably because the bulk of the fluid travels a shorter path than the center line; this agrees with Larrain and Bonilla [23] who used computer algebra to also find high order approximations to the flow in a coiled pipe.

The cross-pipe velocities in a helical pipe are indicated in Figure 2(b) where the torsion induces velocities proportional to those shown in the figure; the generally upward velocity matches the upward twist of positive torsion. Observe that torsion, τ, and variations in curvature, κ', affect only the cross-stream velocities and do not influence the axial velocity u to this order. Conversely, observe that this Stokes flow does *not* have cross-pipe circulation—the strong viscosity eliminates inertia. Instead the curvature of the pipe just skews and alters the velocity field. For curvature $\kappa \neq 0$ the maximum of the axial velocity u increases and moves toward the inner wall of the pipe. This last effect, though seemingly small even for the large curvature of $\kappa = 0.8$ used in Figure 2, is enough to have a strong influence on the shear dispersion as seen in its coefficient (6).

2.3. Laminar flow at finite Reynolds number.

Incorporating the terms representing the advection of fluid momentum into the computer algebra program of Appendix B leads to effects parametrized by the Reynolds number Re. I find that the fluid fields given previously in (17)–(21) are modified by the addition of the following terms:

$$\begin{aligned}
u = \cdots &+ \frac{\mathrm{Re}}{144} r(1-r^2)(29+5r^2-3r^4)\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] \\
&- \frac{\mathrm{Re}^2}{1440} r(1-r^2)(19-21r^2+9r^4-r^6)\cos\theta\,\kappa \\
&+ \frac{\mathrm{Re}^3}{1814400} r(1-r^2)(2969-4381r^2+3249r^4-1301r^6 \\
&\quad + 274r^8-20r^{10})\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] + \mathcal{O}(\kappa^2,\delta^2)\,,
\end{aligned} \tag{21}$$

$$\begin{aligned}
v = \cdots &- \frac{\mathrm{Re}}{72}(1-r^2)^2(4-r^2)\cos\theta\,\kappa \\
&+ \frac{\mathrm{Re}^2}{8640}(1-r^2)^2(13-15r^2+7r^4-r^6)\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] \\
&+ \mathcal{O}(\kappa^2,\delta^2)\,,
\end{aligned} \tag{22}$$

$$\begin{aligned}
w = \cdots &+ \frac{\mathrm{Re}}{72}(1-r^2)(4-23r^2+7r^4)\sin\theta\,\kappa \\
&+ \frac{\mathrm{Re}^2}{8640}(1-r^2)(13-224r^2+266r^4-124r^6 \\
&\quad + 17r^8)\left[-\sin\theta\,\kappa' + \cos\theta\,\kappa\tau\right] + \mathcal{O}(\kappa^2,\delta^2)\,,
\end{aligned} \tag{23}$$

$$\begin{aligned}
p = \cdots &- \frac{\mathrm{Re}}{3} r(9-6r^2+2r^4)\cos\theta\,\kappa \\
&+ \frac{\mathrm{Re}^2}{2160} r(101-120r^2+90r^4-30r^6+3r^8)\left[\cos\theta\,\kappa' + \sin\theta\,\kappa\tau\right] \\
&+ \mathcal{O}(\kappa^2,\delta^2)\,,
\end{aligned} \tag{24}$$

where "$\cdots$" denote the terms already given for Stokes flow in (17)–(21). The modifications to the cross-pipe velocities that are proportional to $\mathrm{Re}\,\kappa (= \mathrm{Dn}^2/4\,\mathrm{Re})$, plotted in Figure 3, agree with those of the Dean flow as used by Johnson and Kamm [19, p. 330] in their work on dispersion. The cross-pipe velocity field exhibits circulation across the pipe induced by the pipe curvature because of fluid inertia. The term in the axial velocity u proportional to $\kappa\,\mathrm{Re}^2 = \mathrm{Dn}^2/4$ also agrees with that of Johnson and Kamm. These terms in the velocity fields are those previously found for the "loosely coiled limit" [3, p. 467] when curvature κ is negligible by itself but the Dean number $\mathrm{Dn} = 2\sqrt{\kappa}\,\mathrm{Re}$ is significant. The above expressions appear to agree precisely with the expressions obtained by Tuttle [39, (58)–(61)] for low Reynolds number flow in a helical pipe; the only differences lie in various factors of two due

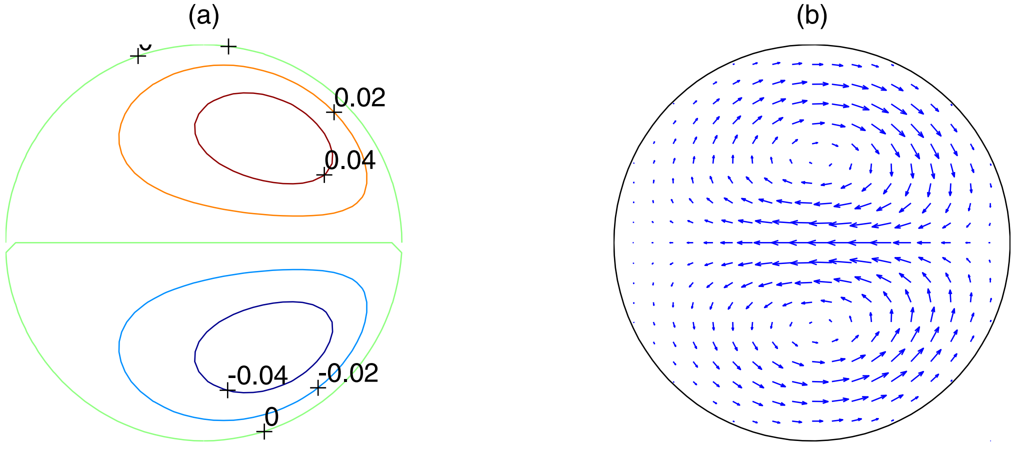


Figure 3. (a) *Axial velocity u contours of the inertial corrections to Poiseuille flow in a helical pipe with curvature $\kappa = 0.5$ to leading order in the torsion τ of the helix.* (b) *Corresponding inertia induced cross-pipe fluid velocities independent of the torsion. The plots are evaluated from an asymptotic solution with errors $\mathcal{O}\big(\kappa^5\big)$.*

to the different nondimensionalization and because my angular velocity w is in a "space-centered" coordinate system, whereas Tuttle's Φ is "body centered."[4] Last, the components of the physical fields (17)–(25) involving $\sin\theta$ and $\cos\theta$ cross-sectional structures agree with those of Hammack and Hydon [14, p. 363], but their formulae have none of the components in $\sin 2\theta$, $\cos 2\theta$ nor a modification to the mean pressure gradient as in (26). Observe in all of the formulae (18)–(25) that torsion coupled to curvature, $\kappa\tau$, appears to have the same effect as longitudinal gradients of curvature, κ', but the fluid fields are rotated in angle by 90°. The above expressions are the first to combine torsion and general variations in curvature.

With fixed fluid flux, the mean pressure gradient, to one higher order in curvature than the fields above, is

$$\begin{aligned}(25)\qquad \bar p' = -8 &+ \left(\tfrac{1}{6} - \tfrac{11}{540}\operatorname{Re}^2 - \tfrac{1541}{32659200}\operatorname{Re}^4\right)\kappa^2 \\ &+ \left(\tfrac{23}{80}\operatorname{Re} + \tfrac{1433}{241920}\operatorname{Re}^3 + \tfrac{6191}{410572800}\operatorname{Re}^5\right)\kappa\kappa' + \mathcal{O}\big(\kappa^3, \delta^2\big)\,.\end{aligned}$$

See from the coefficients plotted in Figure 4 that although the pressure gradient is lessened by curvature for low Reynolds number, for Reynolds number approximately $\operatorname{Re} > 3$ there is an increased pressure gradient loss in a curving pipe. This is attributed to the greater mixing caused by the induced cross-pipe circulation. Also see that a region of tightening curvature, increasing κ^2, has a lesser drop in pressure gradient relative to that for a toroidal pipe, whereas conversely a lessening in the curvature, decreasing κ^2, has a higher drop in the

[4] Although the components of the velocity proportional to curvature κ and the leading order terms in $\operatorname{Re}\kappa'$ reduce to those of Murata, Miyake, and Inaba [27, equations (21)–(22)], in their case of a sinusoidal centerline, the terms in Re^2 are different in detail to those of Murata, Miyake, and Inaba [27, equation (22)], as is the pressure. Terms in curvature κ in the above velocity field agree with those of Pedley [29, equation (4.13)], and terms in the gradient κ' with highest power of Reynolds number Re also agree [29, equations (4.18)–(4.19)] except for the axial velocity u. The above expressions also agree with the small Dean number expansion derived by Kao [21, p. 341] for flow in a helix except for his w_2 which does not match my expression (22) for u.

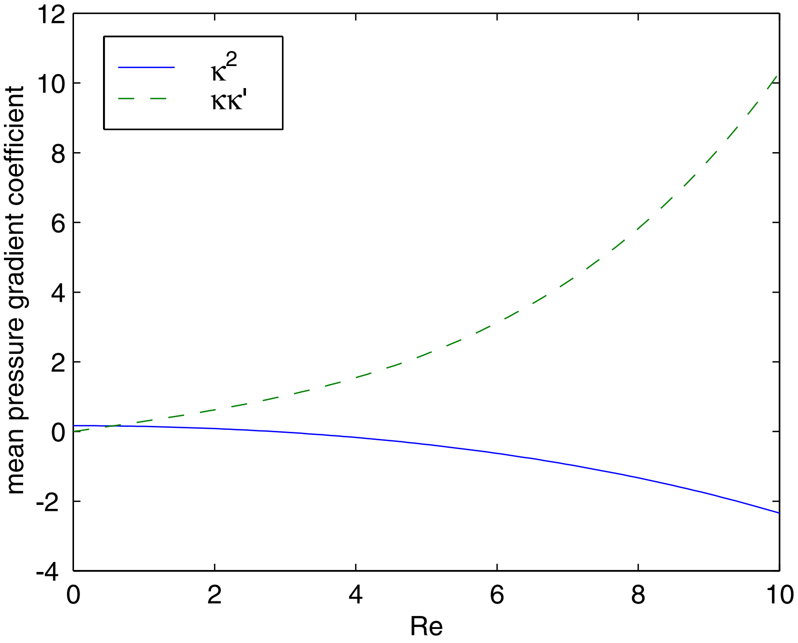


Figure 4. *Coefficients in the expression* (26) *for the mean pressure gradient as a function of the center line curvature* κ *and its gradient* κ'*.*

pressure gradient. This effect is interpreted as a "memory" in the mean pressure gradient of the upstream conditions reflecting the finite distance taken for the flow to relax to the new curvature; retaining just the highest order terms in Re, the mean pressure gradient

$$\begin{aligned} \bar{p}' &\approx -8 - \tfrac{1541}{32659200} \operatorname{Re}^4 \kappa^2 + \tfrac{6191}{821145600} \operatorname{Re}^5 2\kappa\kappa' \\ &\approx -8 - \tfrac{1541}{32659200 \times 16} \operatorname{Dn}^4 \Big|_{s - 4337 \operatorname{Re}/271216}, \end{aligned} \tag{26}$$

where the evaluation of the Dean number at an effective distance upstream of approximately Re /6 pipe radii seems to show the typical distance necessary for the fluid flow to develop in order to accord with the curvature of the pipe. This agrees qualitatively with experiments on a pipe with a finite bend as discussed by Berger, Talbot, and Yao [3, p. 494], where the influence of the bend on the mean pressure gradient extends far downstream. This upstream memory also matches nicely with the commonly quoted distance, $l_s \approx \frac{1}{4} a \operatorname{Re}$ [12] and [3, p. 488], required for flow entering a straight pipe to become fully developed, and with the observation by Murata, Miyake, and Inaba [27, section 4] that the flow in a sinusoidally bent pipe has a lag in its adaptation to the local conditions, the lag increasing with increasing Reynolds number.

The above formulae for both viscous and inertia effects of curvature and torsion are only of low order. Computer algebra computes the velocity field to as high an order as is necessary for the demands of modeling the dispersion in the pipe, described in the next section. Van Dyke [3, p. 475] has shown the asymptotic expansions of the fluid flow field converge for Dean number $\operatorname{Dn} < 96.8/4\sqrt{2} = 17.1$ (for negligible κ but finite Dn). However, I have not explored this issue as here we are primarily concerned with the advection-diffusion model (2) of the dispersion and *its* asymptotic approximations.

3. Advection-dispersion along the curving pipe.

Having determined the fluid flow within the pipe, I now address the advection and longitudinal dispersion within the pipe. We solve the advection-diffusion equation (2) for the evolving concentration $c(s, r, \vartheta, t)$ of contaminant within the fluid. Writing the concentration in terms of the cross-pipe average $C(s, t)$ and its derivatives, I use center manifold techniques (see Appendix A) to construct the Taylor model (4), and its higher order generalizations, of the advection-dispersion in the bent and twisted pipe.

3.1. The concentration field within the pipe.

Using the coordinate system described in section 2.1, the advection-diffusion equation (2) for the evolution of the concentration of the contaminant is

$$\begin{aligned}(27)\qquad &\frac{\partial c}{\partial t} + \mathrm{Pe}\left(\frac{u}{h_s}\frac{\partial c}{\partial s} + v\frac{\partial c}{\partial r} + \frac{w}{r}\frac{\partial c}{\partial \vartheta}\right)\\ &= \frac{1}{rh_s}\left[\frac{\partial}{\partial s}\left(\frac{r}{h_s}\frac{\partial c}{\partial s}\right) + \frac{\partial}{\partial r}\left(rh_s\frac{\partial c}{\partial r}\right) + \frac{\partial}{\partial \vartheta}\left(\frac{h_s}{r}\frac{\partial c}{\partial \vartheta}\right)\right].\end{aligned}$$

This is solved with no flux through the circular walls of the pipe: $\partial c/\partial r = 0$ on $r = 1$.

The computer algebra program (Appendix B) simultaneously determines from the contaminant conservation equation (28) the dynamics on the low-dimensional center manifold, namely, the Taylor model (4). Over a cross-pipe diffusion time the concentration field evolves to be, for example, approximately

$$\begin{aligned}(28)\qquad c = C &- \mathrm{Pe}\frac{\partial C}{\partial s}\frac{1}{24}\left(2 - 6r^2 + 3r^4\right) + \mathrm{Pe}\frac{\partial C}{\partial s}\kappa\cos\theta\left[-\frac{1}{6}\left(4r - 3r^3 + r^5\right)\right.\\ &\quad + \frac{\mathrm{Re}^2}{172800}\left(256r - 285r^3 + 200r^5 - 75r^7 + 15r^9 - r^{11}\right)\\ &\quad \left. + \frac{\mathrm{Sc}\,\mathrm{Re}^2}{34560}\left(68r - 120r^3 + 130r^5 - 75r^7 + 21r^9 - 2r^{11}\right)\right]\\ &+ \mathcal{O}(\kappa^2, \delta^2).\end{aligned}$$

To this order neither torsion nor gradients of curvature affect the concentration field within the pipe, but this is not surprising as one order of δ is counted in $\partial C/\partial s$ leaving no scope for derivatives of the curvature κ to be involved in the above terms. Expressions for the concentration field to the next order in either the curvature κ or longitudinal derivatives δ are algebraically formidable and are not recorded.

From such expressions, the computer algebra also determines the mean flux of contaminant through a cross-section of the pipe:

$$F(C, s) = \overline{uc - \frac{1}{h_s}\frac{\partial c}{\partial s}},$$

where the overbar denotes the average over a cross-section. Then by conservation of contaminant, the model for the evolution of the contaminant is known to follow the conservation equation $\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial s}$. Using the flux F determined above to be correct to some order in axial derivatives, some order in δ, I determine the model of axial dispersion to one order higher in axial derivatives than would otherwise be possible because of the extra axial derivative in the right-hand side of the conservation equation $\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial s}$.

3.2. Arbitrary curvature exhibits upstream memory. The error of $\mathcal{O}(\delta)$ in the shear dispersion coefficient given by (6) encompasses modifications due to both torsion τ and variations along the pipe in the curvature κ. The torsion affects only the dispersion coefficient at $\mathcal{O}(\kappa^2\tau^2)$ and so does not show up in (6).

Variations in curvature along the pipe ($\kappa' \neq 0$) cause the effective dispersion coefficient to become, using $\mathfrak{R}$ to denote the scaled Reynolds number $\operatorname{Re}/10$ and remembering $\operatorname{Pe} = \operatorname{Re}\operatorname{Sc}$,

$$
\begin{aligned}
&D = 1 + \frac{1}{4}\kappa^2 + \frac{7}{48}\left(\kappa\kappa'\right)' - \frac{7}{96}\left(\kappa^2\tau^2 + \kappa'^2\right) \qquad \text{(29a)}\\
&\quad + \frac{\operatorname{Pe}^2}{48}\left[1 + \kappa^2\left(-\tfrac{64225}{171072}\mathfrak{R}^4\operatorname{Sc}^2 + \tfrac{2995}{24192}\mathfrak{R}^4 - \tfrac{36335}{12096}\mathfrak{R}^2 + \tfrac{863}{120}\right)\right] \qquad \text{(29b)}\\
&\quad + \operatorname{Pe}\kappa\kappa'\left[\mathfrak{R}^6\left(\tfrac{9050586625}{26900729856}\operatorname{Sc}^4 + \tfrac{246093875}{1630347264}\operatorname{Sc}^3 - \tfrac{6234774125}{53801459712}\operatorname{Sc}^2\right.\right.\\
&\qquad\quad \left. - \tfrac{1760495125}{40351094784}\operatorname{Sc}\right) \quad + \mathfrak{R}^4\left(-\tfrac{1068925}{2322432}\operatorname{Sc}^3 + \tfrac{33738035}{20901888}\operatorname{Sc}^2 + \tfrac{2310385}{4478976}\operatorname{Sc}\right)\\
&\qquad \left. + \mathfrak{R}^2\left(-\tfrac{383695}{96768}\operatorname{Sc}^2 + \tfrac{19465}{12096}\operatorname{Sc} + \tfrac{985}{12096}\right) \quad - \tfrac{13}{32}\right] \qquad \text{(29c)}\\
&\quad + \operatorname{Pe}^2(\kappa^2\tau^2 + \kappa'^2)\left[\mathfrak{R}^6\left(\tfrac{2542365125}{58692501504}\operatorname{Sc}^4 - \tfrac{1039029345155}{126541033242624}\operatorname{Sc}^2\right.\right.\\
&\qquad\quad \left. - \tfrac{5542735225}{3515028701184}\right) \quad + \mathfrak{R}^4\left(\tfrac{14791164485}{33108590592}\operatorname{Sc}^2 + \tfrac{739414405}{11036196864}\right)\\
&\qquad \left. + \mathfrak{R}^2\left(-\tfrac{7181}{9072}\operatorname{Sc}^2 + \tfrac{7619671}{41803776}\right) \quad - \tfrac{5357}{55296}\right] \qquad \text{(29d)}\\
&\quad + \operatorname{Pe}\mathfrak{R}(\kappa\kappa')'\left[\mathfrak{R}^6\left(-\tfrac{2219783253125}{3012881743872}\operatorname{Sc}^5 - \tfrac{3007270625375}{9038645231616}\operatorname{Sc}^4\right.\right.\\
&\qquad\quad \left. + \tfrac{7670650920025}{63270516621312}\operatorname{Sc}^3 + \tfrac{2879131496575}{23726443732992}\operatorname{Sc}^2 + \tfrac{396673566125}{15817629155328}\operatorname{Sc}\right)\\
&\qquad + \mathfrak{R}^4\left(\tfrac{318184675}{306561024}\operatorname{Sc}^4 - \tfrac{3159503125}{752467968}\operatorname{Sc}^3 - \tfrac{9765145925}{16554295296}\operatorname{Sc}^2 - \tfrac{2663242675}{5518098432}\operatorname{Sc}\right)\\
&\qquad + \mathfrak{R}^2\left(\tfrac{40508065}{4644864}\operatorname{Sc}^3 - \tfrac{3081595}{1548288}\operatorname{Sc}^2 - \tfrac{3844625}{2612736}\operatorname{Sc} - \tfrac{16835}{373248}\right)\\
&\qquad \left. + \tfrac{32413}{27648}\operatorname{Sc} - \tfrac{191}{1728}\right] + \mathcal{O}(\kappa^4, \delta^3)\,. \qquad \text{(29e)}
\end{aligned}
$$

In this large but comprehensive expression observe the following:

Terms (29a) give the effective diffusivity along the center line of the pipe of the molecular diffusion within the bending and twisting geometry of the pipe when there is no flow.

Terms (29b) give the usual shear enhanced dispersion in a straight pipe, $\operatorname{Pe}^2/48$, modified by the leading order (quadratic) effects of pipe curvature. These were the terms of the shear dispersion discussed in the Introduction; see (6). See the curvature induced contribution, the term multiplied by κ^2, plotted relative to the Taylor dispersion coefficient, $1 + \operatorname{Pe}^2/48$, in Figure 5.

Terms (29c) give the leading order effects on the dispersion due to variations in curvature along the pipe; see Figure 6.

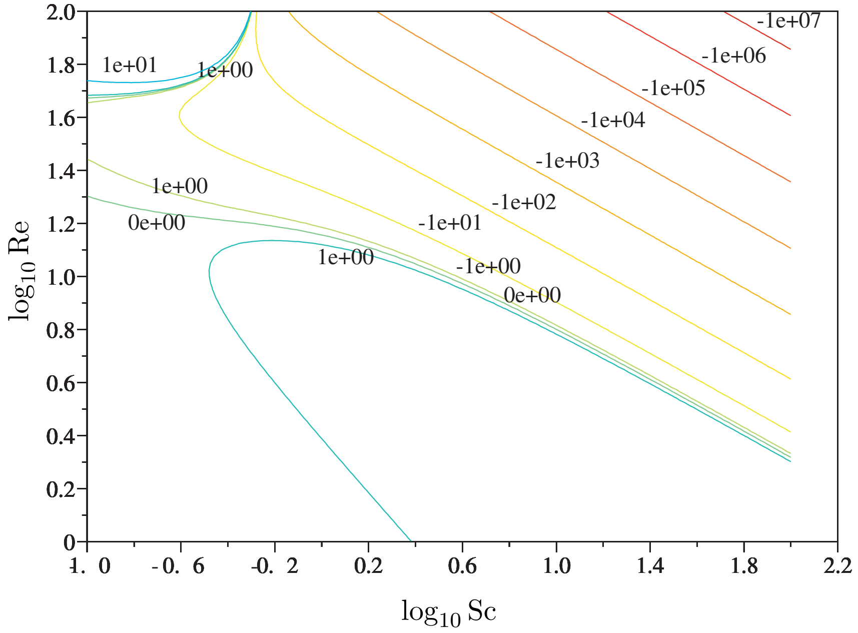


Figure 5. *Here, and analogously in the next three figures, are shown contours of the coefficient of the contribution of plain curvature* κ^2, relative to the *Taylor dispersion coefficient* $1 + \mathrm{Pe}^2\,\mathrm{Sc}^2/48$ *which ranges up to* 10^6 *in the top right corner of the figures, in the dispersion coefficient* (29) *as a function of Reynolds number and Schmidt number. That is, we plot the relative magnitude of* κ^2 *coefficient in the contribution* (29b) *for* $1 < \mathrm{Re} < 100$ *and* $0.1 < \mathrm{Sc} < 100$ *. The contours are the powers of ten* $\pm 10^n$ *for* $n = 0, \ldots, 9$ *, where the red contours are negative, and the blue are positive.*

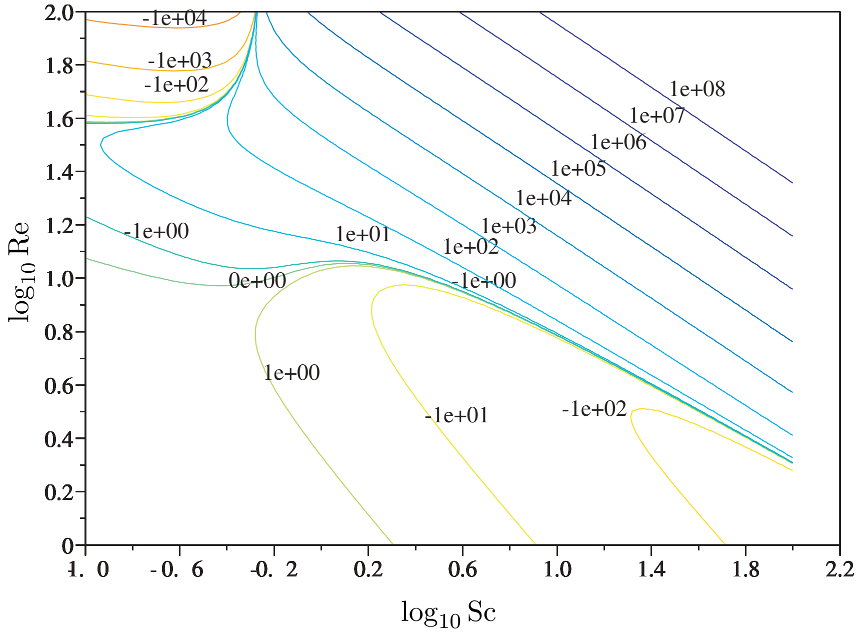


Figure 6. *As in Figure* 5 *but now the contours of the coefficient of* $\kappa\kappa'$ *in* (29c), *relative to the Taylor coefficient* $1 + \mathrm{Pe}^2\,\mathrm{Sc}^2/48$ *, showing the leading order effects due to variations in pipe curvature.*

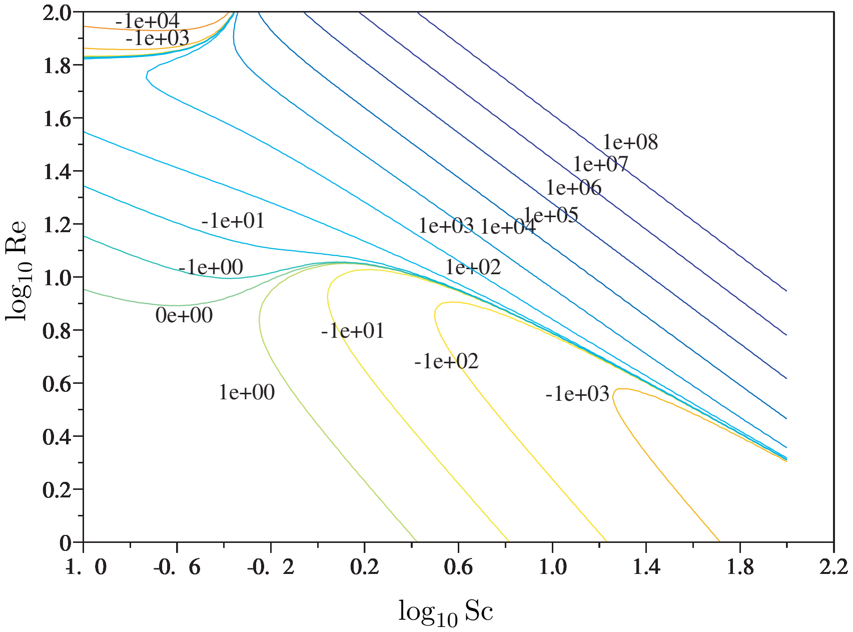


Figure 7. *As in Figure 5 but now the contours of the coefficient of* $(\kappa^2\tau^2 + \kappa'^2)$ *in* (29d)*, relative to the Taylor coefficient* $1 + \mathrm{Pe}^2\,\mathrm{Sc}^2/48$*, showing the leading order effects on dispersion due to pipe torsion.*

Terms (29d) give the leading order effect of torsion on the dispersion, namely, quadratic but moderated by the multiplication by κ^2; see Figure 7;

Terms (29d)–(29e) through κ'^2 and κ'' give second order effects of the variations in curvature; see Figures 7 and 8.

It is intriguing to see that the effects of torsion and second order gradients of curvature factorize as shown in (29d)–(29e). I suggest the reason for this factorization is due to two effects: first, upstream "memory" of the dispersion, to be discussed later, involves

$$\kappa^2\Big|_{s-\xi} = \kappa^2 - 2\xi\kappa\kappa' + \xi^2(\kappa\kappa')' + \mathcal{O}(\xi^3),$$

which may explain the appearance of the combination $(\kappa\kappa')'$; and second, curvature gradients and torsion, κ' and $\kappa\tau$, respectively, both create the same but orthogonal structures in the fluid flow as commented after (22)–(25).

For large Schmidt number Sc *(typical for material dispersion in liquids).* There are two distinguished limits of the above expression for the effective dispersion coefficient, the second being a subset of the first.

- First, for large Schmidt number Sc the highest powers of Sc dominate. However, in various subexpressions they appear in combination with the Reynolds number $\mathrm{Re} = 10\mathfrak{R}$. Thus there is a distinguished limit with large Sc and small Re in which $\mathrm{Re}^2\,\mathrm{Sc}$ is of order 1 (near the lower right region of Figures 5–8. In terms of the magnitude δ of the slow axial variations, an appropriate scaling is that the fluid flow is slow, $\mathrm{Re} \sim \delta$, and the Schmidt number is large enough, $\mathrm{Sc} \sim 1/\delta^2$, so that the Peclet number is also large, $\mathrm{Pe} \sim 1/\delta$; then the effective diffusion coefficient is large, $D \sim 1/\delta^2$. Using these

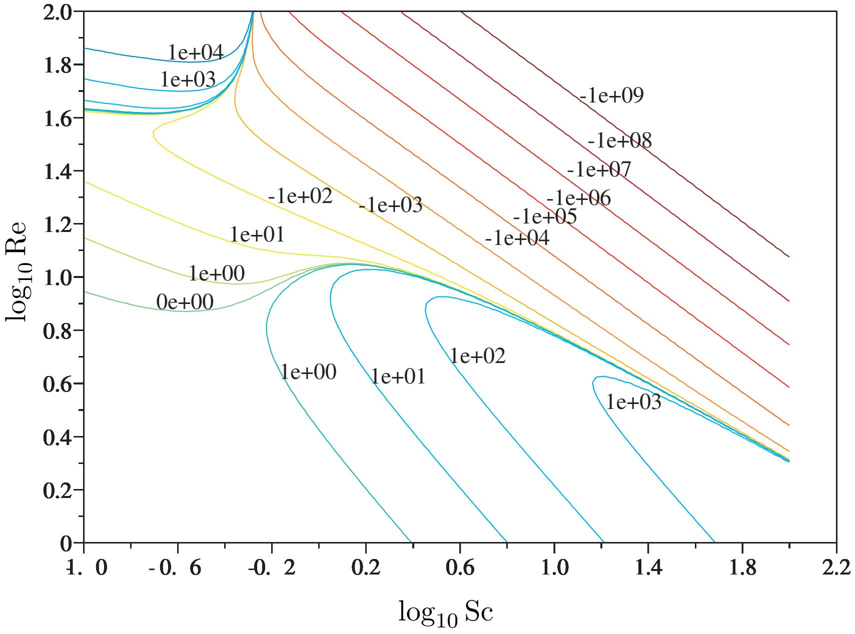


Figure 8. *As in Figure 5 but now the contours of the coefficient of* $(\kappa\kappa')'$ *in* (29e)*, relative to the Taylor coefficient* $1 + \mathrm{Pe}^2\,\mathrm{Sc}^2/48$*, showing second order effects of pipe curvature.*

orders of magnitude, introducing the order 1 parameter $\alpha = \mathrm{Re}\,\mathrm{Pe}/100 = \mathrm{Re}^2\,\mathrm{Sc}/100$ and evaluating fractions, the leading order terms in the dispersion coefficient (29) are

$$\begin{aligned} D \approx{}& .02083\,\mathrm{Pe}^2 + (.1498 - .007821\,\alpha^2)\,\mathrm{Pe}^2\,\kappa^2 \\ &+ (-.03965 - .004603\,\alpha + .003364\,\alpha^2)\,\mathrm{Pe}^3\,\kappa\kappa' \\ &+ (-.007916 + .0004332\,\alpha^2)\,\mathrm{Pe}^4(\kappa^2\tau^2 + \kappa'^2) \\ &+ (.008721 + .001038\,\alpha - .0007368\,\alpha^2)\,\mathrm{Pe}^4(\kappa\kappa')'\,. \end{aligned} \tag{30}$$

- Second, for a typical Schmidt number Sc bigger than 10^3 or so, and for any flow with Reynolds number Re bigger than about 1, the parameter α will be bigger than about 10 and the above expression (31) will be dominated by the quadratic powers in α; this is shown by the near linear contours in the upper right region of Figures 5–8. That is, the dispersion coefficient

$$\begin{aligned} D \approx \left(\frac{\mathrm{Pe}}{10}\right)^2 \Bigg\{ 2.083 + \left(\frac{\mathrm{Re}^2\,\mathrm{Sc}}{100}\right)^2 \Bigg[& -0.7821\,\kappa^2 + 3.364\,\frac{\mathrm{Pe}}{10}\kappa\kappa' \\ & + \left(\frac{\mathrm{Pe}}{10}\right)^2 \Big(4.332\,(\kappa^2\tau^2 + \kappa'^2) - 7.368\,(\kappa\kappa')'\Big) \Bigg] \Bigg\}. \end{aligned} \tag{31}$$

We noted in (27) that the mean pressure gradient in the fluid flow at any location was appropriate to the curvature some distance upstream. Similar memory effects are seen in the dispersion coefficient again due to the finite time taken for the flow and the dispersion to

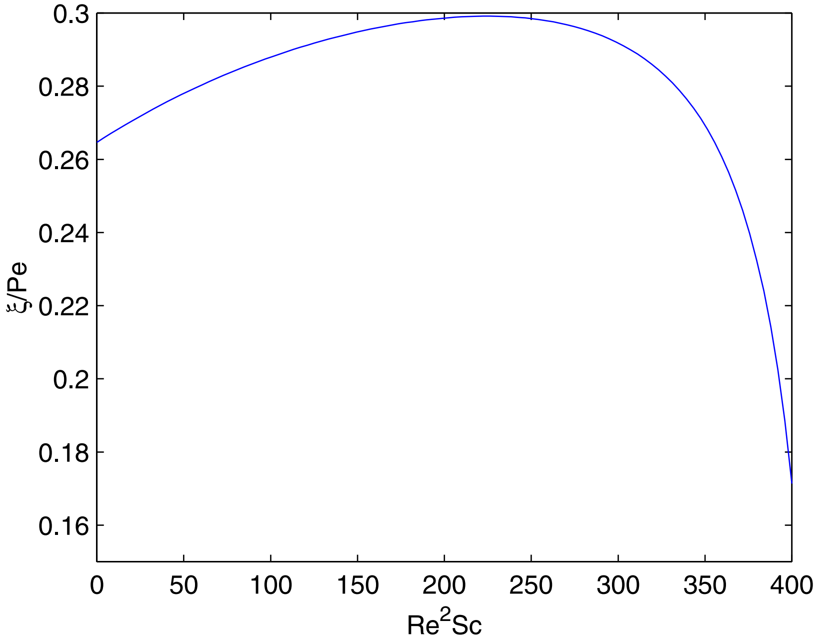


Figure 9. *In a generally curving pipe the effective dispersion has the value appropriate to the curvature a distance ξ upstream.*

relax to new curvature. The subexpression $-0.7821\,\kappa^2 + 0.3364\,\mathrm{Pe}\,\kappa\kappa'$ appearing in the first line of (31) is equivalent to simply $-0.7821\,\kappa^2$ evaluated at a distance $\xi = 0.2151\,\mathrm{Pe}$ upstream from any particular location. I do not attempt to complicate this memory effect any further by trying to include the second order term $(\kappa\kappa')'$, as there is a plethora of possibilities, but for the purposes of discussion I assume both $\kappa\kappa'$ and $(\kappa\kappa')'$ terms are attributable to upstream memory. The ratio of the coefficients of $\kappa\kappa'$ and κ^2 in (31), shown in Figures 6 and 5, respectively, similarly quantify the upstream memory for low Reynolds number flows as shown in Figure 9. Such memory effects in shear dispersion in varying channels were first recognized by Smith [34].

From Figure 7 and the coefficient approximations (31) and (31), see that torsion and curvature gradients generally enhance dispersion along the pipe except for low Reynolds numbers, $\mathrm{Re}^2\,\mathrm{Sc} < 427.4$, when they make the dispersion coefficient smaller. However, the effect torsion has upon the dispersion coefficient seems small because not only is the effect quadratic in the torsion τ, but it is also ameliorated by the multiplication by the curvature squared. However, ignoring the $\kappa\kappa'$ terms and noting that $\kappa' = \kappa(\log\kappa)'$, I write (31) as

$$\begin{aligned} D \approx \left(\frac{\mathrm{Pe}}{10}\right)^2 \Bigg\{ & 2.083 \\ & - 0.7821 \left(\frac{\kappa\,\mathrm{Re}^2\,\mathrm{Sc}}{100}\right)^2 \left[1 - \left(\frac{\mathrm{Pe}}{10}\right)^2 5.539\,(\tau^2 + (\log\kappa)'^2)\right] \Bigg\}. \end{aligned} \tag{32}$$

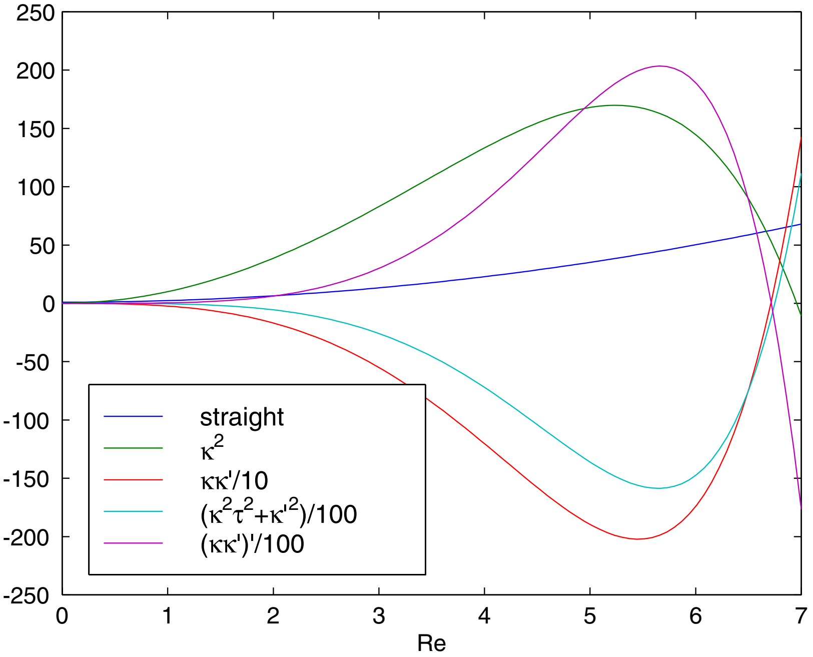


Figure 10. *Coefficients of the dispersion coefficient D for water at $15^\circ C$ given by* (34)*. For higher Reynolds numbers the highest powers in* Re *dominate the coefficients.*

This suggests that torsion, or proportional gradients of curvature, greater than about $4/\operatorname{Pe}$ may cause the dispersion coefficient D to increase with curvature κ, instead of decreasing. That torsion could eliminate the increased mixing due to secondary circulations seems unlikely, so I predict higher order terms in the torsion τ would limit its influence on the dispersion.

For the dispersion of heat in water. with a Prandtl number of $\operatorname{Sc} = 8.1$ at 15°C ($\log_{10}\operatorname{Sc} = 0.91$ in Figures 5–8) the dispersion coefficient (29) reduces to

$$\begin{aligned} D &= 1.3669\operatorname{Re}^2 + 1 \\ &+ \kappa^2\left(-.003350\operatorname{Re}^6 - .04106\operatorname{Re}^4 + 9.830\operatorname{Re}^2 + .25\right) \\ &+ \kappa\kappa'\left(.01232\operatorname{Re}^7 - .1090\operatorname{Re}^5 - 2.01\operatorname{Re}^3 - 3.291\operatorname{Re}\right) \\ &+ (\kappa^2\tau^2 + \kappa'^2)\left(.01220\operatorname{Re}^8 + .1928\operatorname{Re}^6 - 33.95\operatorname{Re}^4 - 6.356\operatorname{Re}^2 - .07292\right) \\ &+ (\kappa\kappa')'\left(-.02191\operatorname{Re}^8 + .1777\operatorname{Re}^6 + 36.39\operatorname{Re}^4 + 7.602\operatorname{Re}^2 + .1458\right). \end{aligned} \tag{33}$$

For Reynolds number $\operatorname{Re} > 7$ the highest powers in Re dominate; see Figure 10 for the dependence on smaller Re. Observe that for Reynolds number $\operatorname{Re} > 6.95$ curvature enhances the dispersion of heat and vice-versa, whereas for $\operatorname{Re} > 6.74$ torsion and curvature gradients reduce the dispersion and vice-versa.

For the dispersion of heat in air. with a Prandtl number of $\operatorname{Sc} = 0.71$ at 15°C ($\log_{10}\operatorname{Sc} = -0.15$ in Figures 5–8) and recalling that $\mathfrak{R} = \operatorname{Re}/10$, the dispersion coefficient (29) reduces to

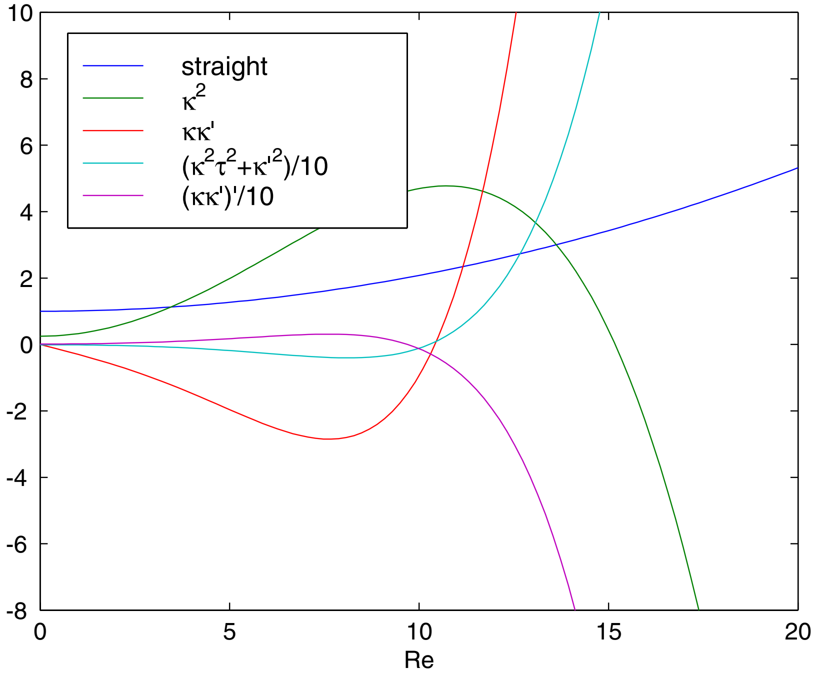


Figure 11. *Coefficients of the dispersion coefficient D for water at $15^\circ C$ given by (35). For higher Reynolds numbers the highest powers in* Re *dominate the coefficients.*

$$
\begin{aligned}
(34)\qquad D &= 1.080\,\mathfrak{R}^2 + 1 \\
&+ \kappa^2\left(-.07649\,\mathfrak{R}^6 - 3.244\,\mathfrak{R}^4 + 7.767\,\mathfrak{R}^2 + .25\right) \\
&+ \kappa\kappa'\left(.3979\,\mathfrak{R}^7 + 7.462\,\mathfrak{R}^5 - 5.871\,\mathfrak{R}^3 - 2.925\,\mathfrak{R}\right) \\
&+ (\kappa^2\tau^2 + \kappa'^2)\left(.3011\,\mathfrak{R}^8 + 15.48\,\mathfrak{R}^6 - 11.82\,\mathfrak{R}^4 - 5.022\,\mathfrak{R}^2 - .07292\right) \\
&+ (\kappa\kappa')'\left(-.7615\,\mathfrak{R}^8 - 13.98\,\mathfrak{R}^6 + 8.055\,\mathfrak{R}^4 + 5.282\,\mathfrak{R}^2 + .1458\right)\,.
\end{aligned}
$$

For Reynolds number $\mathrm{Re} > 15$ the highest powers in Re dominate; see Figure 11 for the dependence on smaller Re. Observe that for Reynolds number $\mathrm{Re} > 15.18$ curvature enhances the dispersion of heat and vice-versa, whereas for $\mathrm{Re} > 10.35$ torsion and curvature gradients reduce the dispersion and vice-versa.

3.3. Skewness is very sensitive to curvature. Computer algebra straightforwardly determines high order terms in the advection-diffusion equation (2). Chatwin [5] investigated the relatively slow approach to normality in shear dispersion. However, variations in pipe curvature and torsion distort any normal profile. Hence expect such variations to have a large effect on skewness.

The third order modification to the Taylor model of dispersion is

$$
(35)\qquad \frac{\partial C}{\partial t} \approx -\mathrm{Pe}\frac{\partial C}{\partial s} + \frac{\partial}{\partial s}\left(D\frac{\partial C}{\partial s}\right) + \frac{\partial}{\partial s}\left(E\frac{\partial^2 C}{\partial s^2}\right),
$$

where the skewness coefficient

$$\begin{align}
E = {}&-\frac{\mathrm{Pe}^3}{2880} \tag{36a}\\
&+ \mathrm{Pe}\,\kappa^2\Big[\mathfrak{R}^6\Big(\tfrac{3241338875}{107602919424}\,\mathrm{Sc}^4+\tfrac{1104359125}{35867639808}\,\mathrm{Sc}^2\Big)-\tfrac{4085615}{20901888}\mathfrak{R}^4\,\mathrm{Sc}^2 \notag\\
&\quad+\mathfrak{R}^2\Big(\tfrac{68855}{96768}\,\mathrm{Sc}^2+\tfrac{985}{12096}\Big)-\tfrac{13}{32}\Big] \tag{36b}\\
&+\kappa\kappa'\Big[\mathfrak{R}^8\Big(-\tfrac{5943982203125}{9038645231616}\,\mathrm{Sc}^6-\tfrac{1183708683125}{2259661307904}\,\mathrm{Sc}^5-\tfrac{12012557065375}{9038645231616}\,\mathrm{Sc}^4 \notag\\
&\qquad-\tfrac{4304114889625}{5931610933248}\,\mathrm{Sc}^3\Big) \notag\\
&\quad+\mathfrak{R}^6\Big(\tfrac{1250362375}{1379524608}\,\mathrm{Sc}^5+\tfrac{820657375}{64665216}\,\mathrm{Sc}^4+\tfrac{14670875}{4138573824}\,\mathrm{Sc}^3+\tfrac{1214048125}{4138573824}\,\mathrm{Sc}^2\Big) \notag\\
&\quad+\mathfrak{R}^4\Big(-\tfrac{11414525}{290304}\,\mathrm{Sc}^4+\tfrac{1917625}{193536}\,\mathrm{Sc}^3-\tfrac{25984475}{5225472}\,\mathrm{Sc}^2-\tfrac{84175}{93312}\,\mathrm{Sc}\Big) \notag\\
&\quad+\mathfrak{R}^2\Big(\tfrac{17795}{1152}\,\mathrm{Sc}^2-\tfrac{955}{432}\,\mathrm{Sc}\Big)+\tfrac{7}{12}\Big] \tag{36c}\\
&+\mathrm{Pe}(\kappa^2\tau^2+\kappa'^2)\Big[\mathfrak{R}^8\Big(\tfrac{254564364353125}{1434131710083072}\,\mathrm{Sc}^6-\tfrac{9130311425570375}{34419161041993728}\,\mathrm{Sc}^4 \notag\\
&\qquad-\tfrac{545957180588375}{29041167129182208}\,\mathrm{Sc}^2\Big) \notag\\
&\quad+\mathfrak{R}^6\Big(\tfrac{9773515705025}{2259661307904}\,\mathrm{Sc}^4+\tfrac{241252377733475}{759246199455744}\,\mathrm{Sc}^2\Big) \notag\\
&\quad+\mathfrak{R}^4\Big(-\tfrac{1227346555}{114960384}\,\mathrm{Sc}^4-\tfrac{24438946685}{16554295296}\,\mathrm{Sc}^2-\tfrac{159535}{2128896}\Big) \notag\\
&\quad+\mathfrak{R}^2\Big(\tfrac{1325215}{258048}\,\mathrm{Sc}^2-\tfrac{118243}{580608}\Big)+\tfrac{801}{2560}\Big] \tag{36d}\\
&+\mathrm{Pe}(\kappa\kappa')'\Big[\mathfrak{R}^8\Big(-\tfrac{26098554271144375}{206514966251962368}\,\mathrm{Sc}^6+\tfrac{10015284615625}{204875958583296}\,\mathrm{Sc}^5 \notag\\
&\qquad+\tfrac{83884205830882375}{206514966251962368}\,\mathrm{Sc}^4+\tfrac{68444215116292625}{4646586740669153328}\,\mathrm{Sc}^3+\tfrac{44871973001125}{691456360218624}\,\mathrm{Sc}^2\Big) \notag\\
&\quad+\mathfrak{R}^6\Big(\tfrac{71359651375}{6025763487744}\,\mathrm{Sc}^5-\tfrac{516258977875}{111588212736}\,\mathrm{Sc}^4-\tfrac{596762793925}{2875932573696}\,\mathrm{Sc}^3 \notag\\
&\qquad-\tfrac{2958969842375}{23007460589568}\,\mathrm{Sc}^2-\tfrac{62332574575}{1977203644416}\,\mathrm{Sc}\Big) \notag\\
&\quad+\mathfrak{R}^4\Big(\tfrac{48287105}{3784704}\,\mathrm{Sc}^4-\tfrac{480287515}{172440576}\,\mathrm{Sc}^3+\tfrac{4233379915}{2759049216}\,\mathrm{Sc}^2+\tfrac{89672375}{459841536}\,\mathrm{Sc}+\tfrac{159535}{2128896}\Big) \notag\\
&\quad+\mathfrak{R}^2\Big(-\tfrac{23023685}{4644864}\,\mathrm{Sc}^2+\tfrac{2964127}{4644864}\,\mathrm{Sc}+\tfrac{501107}{2322432}\Big)-\tfrac{2257}{5760}\Big]+\mathcal{O}(\kappa^4,\delta^3)\,. \tag{36e}
\end{align}$$

The skewness coefficient for flow in a straight pipe, $-\mathrm{Pe}^3/2880$ from (36a), is well known [6]. One outstanding puzzle in the field of dispersion is that in rivers one observes contaminant concentrations with long tails upstream (not downstream) [2]. But theoretical models predict only either a weak enhancement of upstream tails or, more confoundedly, as the above negative skewness coefficient implies for straight pipe flow, a weak downstream tail. However, as we now see, curvature effects, presumably induced by the secondary flows, greatly change the skewness coefficient thereby enhancing the upstream tail of a contaminant release. With the caveat that this derivation is for laminar flow in pipes and not turbulent flow in rivers, this qualitative match in the theoretical model compared with observations is pleasing.

For large Schmidt number Sc. Typical for the dispersion of material in liquids, and similar to the dispersion coefficient D, there are two distinguished limits of the above expression for the skewness coefficient.

- First, in terms of the magnitude δ of the slow axial variations, the appropriate scaling is that the fluid flow is slow, $\mathrm{Re} \sim \delta$, and the Schmidt number is large enough, $\mathrm{Sc} \sim 1/\delta^2$, so that the Peclet number is also large, $\mathrm{Pe} \sim 1/\delta$; then the skewness coefficient is large, $E \sim 1/\delta^3$. Recalling the parameter $\alpha = \mathrm{Re}\,\mathrm{Pe}\,/100 = \mathrm{Re}^2\,\mathrm{Sc}\,/100$ and evaluating fractions, the leading order terms in the skewness coefficient (36) are

$$\begin{aligned} E \approx -\frac{\mathrm{Pe}^3}{2880}\Big[1 &- \kappa^2\left(.8675\,\alpha^2 + 20.49\right) \\ &+ \mathrm{Pe}\,\kappa\kappa'\left(.1894\,\alpha^2 - .2610\,\alpha + 11.32\right) \\ &+ \mathrm{Pe}^2(\kappa^2\tau^2 + \kappa'^2)\left(-.05112\,\alpha^2 + 3.075\right) \\ &+ \mathrm{Pe}^2(\kappa\kappa')'\left(.03640\,\alpha^2 - .003411\,\alpha - 3.674\right)\Big] . \end{aligned} \tag{37}$$

- Second, for a typical Schmidt number Sc bigger than 10^3 or so, and for any flow with Reynolds number Re bigger than about 2, the parameter α will be bigger than about 40 and the above skewness coefficient (38) is dominated by the quadratic powers in α:

$$\begin{aligned} E \approx \frac{\mathrm{Pe}^3}{2880}\Bigg\{-1 + \left(\frac{\mathrm{Re}^2\,\mathrm{Sc}}{100}\right)^2\Bigg[&0.8675\,\kappa^2 - 1.894\,\frac{\mathrm{Pe}}{10}\kappa\kappa' \\ &+ \left(\frac{\mathrm{Pe}}{10}\right)^2\left(5.112\,(\kappa^2\tau^2 + \kappa'^2) - 3.640\,(\kappa\kappa')'\right)\Bigg]\Bigg\} . \end{aligned} \tag{38}$$

  In this regime, even small curvature, through the κ^2 term, will cause the skewness coefficient to become positive, possibly large, and so will lead to concentration tails upstream (qualitatively as observed in rivers). Torsion in the pipe leads to the same upstream tails.

Recognize another upstream memory effect. The subexpression $0.8675\,\kappa^2 - 0.1894\,\mathrm{Pe}\,\kappa\kappa'$ appearing in the first line of (39) is equivalent to simply $0.8675\,\kappa^2$ evaluated at a distance $\xi = 0.1092\,\mathrm{Pe}$ upstream from any particular location. This upstream memory is approximately half that of the dispersion coefficient.

For the dispersion of heat in water. With a Prandtl number of $\mathrm{Sc} = 8.1$ at 15°C, the skewness coefficient (36) reduces to

$$\begin{aligned} E = {}&-0.1845\,\mathrm{Re}^3 \\ &+ \kappa^2\left(-3.291\,\mathrm{Re} + 3.788\,\mathrm{Re}^3 - .01039\,\mathrm{Re}^5 + .001067\,\mathrm{Re}^7\right) \\ &+ \kappa\kappa'\left(.5833 + 9.956\,\mathrm{Re}^2 - 16.43\,\mathrm{Re}^4 + .08625\,\mathrm{Re}^6 - .002101\,\mathrm{Re}^8\right) \\ &+ (\kappa^2\tau^2 + \kappa'^2)\left(2.534\,\mathrm{Re} + 27.28\,\mathrm{Re}^3 - 37.30\,\mathrm{Re}^5 + .1509\,\mathrm{Re}^7 + .003968\,\mathrm{Re}^9\right) \\ &+ (\kappa\kappa')'\left(-3.174\,\mathrm{Re} - 25.91\,\mathrm{Re}^3 + 43.37\,\mathrm{Re}^5 - .1589\,\mathrm{Re}^7 - .002604\,\mathrm{Re}^9\right) . \end{aligned} \tag{39}$$

For Reynolds number $\mathrm{Re} > 11$ the highest powers in Re dominate. Observe that for these Reynolds numbers both curvature and torsion may easily reverse the sign of the skewness paramater E through the combination

$$+\operatorname{Re}^7\left[.001067\,\kappa^2+.003968\operatorname{Re}^2(\kappa^2\tau^2+\kappa'^2)\right].$$

Again this effect promotes upstream tails in the dispersion. Although the terms in $\kappa\kappa'$ and $(\kappa\kappa')'$ may keep E negative, we prefer to interpret these as representing upstream memory.

For the dispersion of heat in air. With a Prandtl number of $\mathrm{Sc}=0.71$ at 15°C and recalling that $\mathfrak{R}=\operatorname{Re}/10$, the dispersion coefficient (36) reduces to

$$\begin{aligned}E={}&-0.1242\,\mathfrak{R}^3\\&+\kappa^2\left(-2.884\,\mathfrak{R}+3.124\,\mathfrak{R}^3-0.6995\,\mathfrak{R}^5+0.1645\,\mathfrak{R}^7\right)\\&+\kappa\kappa'\left(0.5833+6.217\,\mathfrak{R}^2-9.592\,\mathfrak{R}^4+3.537\,\mathfrak{R}^6-0.7762\,\mathfrak{R}^8\right)\\&+(\kappa^2\tau^2+\kappa'^2)\left(2.221\,\mathfrak{R}+16.93\,\mathfrak{R}^3-25.07\,\mathfrak{R}^5+8.940\,\mathfrak{R}^7-0.3844\,\mathfrak{R}^9\right)\\&+(\kappa\kappa')'\left(-2.782\,\mathfrak{R}-12.99\,\mathfrak{R}^3+22.94\,\mathfrak{R}^5-9.478\,\mathfrak{R}^7+1.287\,\mathfrak{R}^9\right).\end{aligned}\tag{40}$$

For Reynolds number $\operatorname{Re}>40$ the highest powers in Re dominate. Observe that for such a larger Reynolds number the sign of the skewness coefficient changes sign sensitively depending upon the torsion τ, curvature κ, and its gradients.

3.4. Higher order curvature affects the dispersion.

Computer algebra also straightforwardly determines even higher order corrections to the dispersion coefficient. These terms may be used, for example, to give estimates of the errors in the earlier approximations. However, the algebraic expressions quickly become extremely complicated. We just extend the analysis to the next order in curvature, but no higher order in gradients, to obtain the following correction to the dispersion coefficient (29):

$$\begin{aligned}D={}&\cdots+\tfrac{1}{8}\kappa^4\\&+\frac{\mathrm{Pe}^2\,\kappa^4}{48}\Big[\operatorname{Re}^8\Big(\tfrac{6959456407}{30946298639155200000}\,\mathrm{Sc}^4+\tfrac{148720297230839}{46465867406691532800000}\,\mathrm{Sc}^2\\&\qquad-\tfrac{11319036743801}{1429718997128970240000}\Big)\\&\quad+\operatorname{Re}^6\Big(\tfrac{21839753491553}{126541033242624000000}\,\mathrm{Sc}^2+\tfrac{1800408289399}{27115935694848000000}\Big)\\&\quad+\operatorname{Re}^4\Big(-\tfrac{24648813997}{64377815040000}\,\mathrm{Sc}^2+\tfrac{5096950451}{21459271680000}\Big)\\&\quad-\tfrac{4685593}{348364800}\operatorname{Re}^2+\tfrac{13829}{9216}\Big]+\mathcal{O}(\kappa^6,\delta)\end{aligned}\tag{41}$$

or approximately

$$\begin{aligned}D\approx{}&\cdots+\tfrac{1}{8}\kappa^4\\&+\frac{\mathrm{Pe}^2\,\kappa^4}{48}\left[\left(\frac{\operatorname{Re}}{10}\right)^8\left(0.22\,\mathrm{Sc}^4+0.032\,\mathrm{Sc}^2-0.079\right)\right.\\&\quad+\left(\frac{\operatorname{Re}}{10}\right)^6\left(1.7\,\mathrm{Sc}^2+0.66\right)+\left(\frac{\operatorname{Re}}{10}\right)^4\left(-3.8\,\mathrm{Sc}^2+2.4\right)\\&\quad\left.-1.3\left(\frac{\operatorname{Re}}{10}\right)^2+1.5\right]+\mathcal{O}(\kappa^6,\delta).\end{aligned}\tag{42}$$

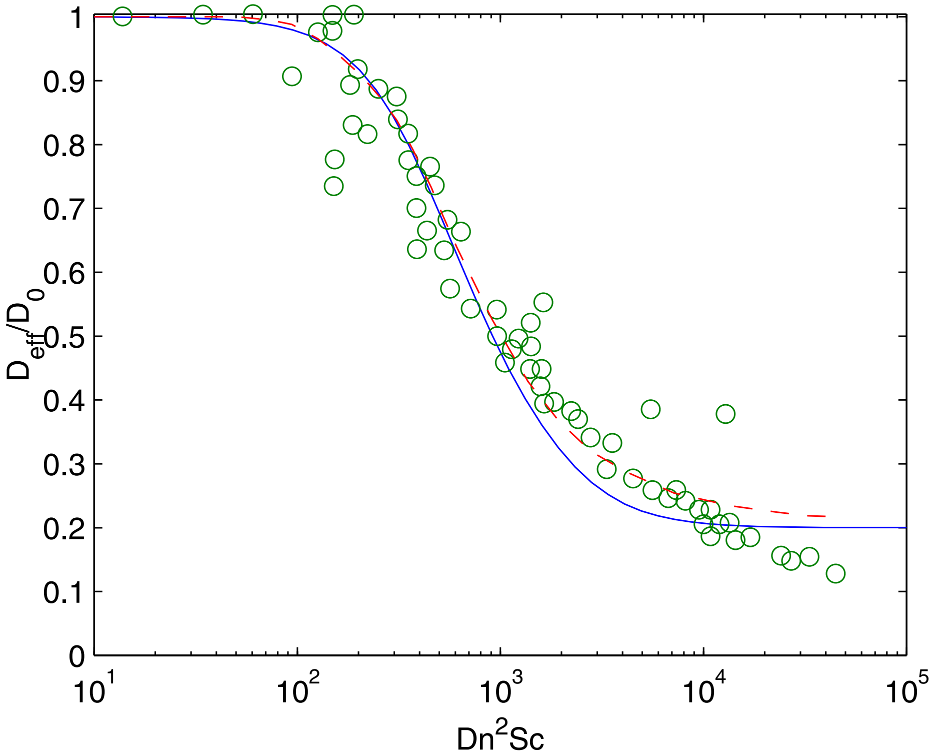


Figure 12. *Comparison of the Padé approximant* (43) *(solid) with experimental estimates (circles) collated by Johnson and Kamm* [19]*, Figure* 9*, and the predictions (dashed line) of their spectral method based on Poiseuille flow.*

I also computed the dispersion coefficient to the next correction, with errors $\mathcal{O}(\kappa^8, \delta)$. Then recalling the Dean number $\mathrm{Dn} = 2\sqrt{\kappa}\,\mathrm{Re}$, the dominant terms for the coefficient of dispersion of material in liquids, large Schmidt number Sc, are

$$D \approx \frac{\mathrm{Pe}^2}{48}\left[1 - 0.3754\left(\frac{\mathrm{Dn}^2\,\mathrm{Sc}}{400}\right)^2 + 0.2249\left(\frac{\mathrm{Dn}^2\,\mathrm{Sc}}{400}\right)^4 \right.$$
$$\left. - 0.1388\left(\frac{\mathrm{Dn}^2\,\mathrm{Sc}}{400}\right)^6 + \mathcal{O}(\,\mathrm{Dn}^{16}\,\mathrm{Sc}^8\,)\right].$$

This expression is valid for small enough $\mathrm{Dn}^2\,\mathrm{Sc}$. Using the additional information [19, section 4.3] that the limit at large $\mathrm{Dn}^2\,\mathrm{Sc}$ is approximately 0.20, I construct the following Padé approximant in terms of $\alpha = \mathrm{Dn}^2\,\mathrm{Sc}/400$:

$$D \approx \frac{\mathrm{Pe}^2}{48} \times \frac{1 + 0.3068\,\alpha^2 + 0.007811\,\alpha^4}{1 + 0.6822\,\alpha^2 + 0.03905\,\alpha^4}\,. \tag{43}$$

See in Figure 12 that this expression for the dispersion coefficient matches reasonably well with experiments over the whole range of $\mathrm{Dn}^2\,\mathrm{Sc}$.

4. Conclusion.

Computer algebra handles the considerable details of deriving the complicated expressions describing dispersion in generally curving pipes (see Appendix B). Fixing one of the fluid flux or the mean pressure gradients affects the dispersion (see section 1), and throughout I present results for the appropriate case of fixed fluid flux. The Padé approximation (43) for the dispersion in a constant curvature pipe is reasonably accurate over the entire range of Dean numbers. When the pipe's curvature and torsion vary, much of the effects of variations upon the dispersion may be recast as an upstream memory. Overall, torsion τ in the pipe seems to have little effect on the dynamics except for the sensitivity, in combination with the curvature κ, of the skewness (see section 3.3). The skewness coefficient is very sensitive to curvature and hence is easily made positive, which may thus qualitatively explain the observations of long upstream tails in the dispersion of material in rivers, despite the differences in detail between turbulent river flow and laminar pipe flow analyzed here.

Appendix A. Center manifold theory constructs the model.

Here we explore in more detail the construction of the model of the dispersion and how it is supported by center manifold theory. This section is based upon arguments for applying center manifold theory in slowly varying long-wave asymptotics [30] and upon later developed straightforward iterative techniques for constructing the model [31].

Consider the advection-diffusion equation (2) for the evolution of the concentration field $c(s, r, \vartheta, t)$. Write it in the form of a dynamical system

$$\frac{\partial c}{\partial t} = \mathcal{L}c + f(c, \epsilon)\,, \tag{44}$$

where $\mathcal{L}$ is a linear operator of cross-pipe diffusion (the r and ϑ operators on the right-hand side of (28)) whose spectrum, as required by center manifold theory, is discrete and separates into zero eigenvalues, the critical ones, and eigenvalues with strictly negative real-part; where ϵ is a vector of parameters representing the pipe's curvature κ and the strength of along pipe gradients δ; and where f is strictly nonlinear when considered as a function of concentration c and parameters ϵ together; that is, in the dispersion all the along pipe processes, including those induced by the geometric variations in the pipe's shape, are placed in f. The critical eigenvalues of $\mathcal{L}$ correspond to the conservation of material in a cross-section when there are no axial variations; the other eigenvalues of $\mathcal{L}$ are all negative (corresponding to a cross-pipe diffusion time). Thus center manifold theory [4] asserts that the concentration field exponentially quickly settles onto a state, approximately (29), parametrized by the cross-sectional average concentration C. The aim is to find a low-dimensional model for the evolution of the cross-sectional average concentration $\frac{\partial C}{\partial t} = g(C)$, such as (4). This model will comprehensively describe all the dynamics after the exponential transients have decayed.

In this application there is one critical mode, associated with the eigenvalue zero, at each of an "infinite" number of cross-sections; thus there is an infinite number of critical modes, parametrized by the axial location s, and so we seek a model expressed in terms of the spatio-temporal function $C(s, t)$. The theoretical support for such infinite dimensional center manifolds [13] is considerably weaker than that for finite dimensional manifolds, but rational arguments support the application of the techniques [30].[5] The critical mode at each cross-section is simply one of constant concentration; thus a "linear" approximation to the center manifold and the evolution thereon is simply

$$c(r,s,\vartheta,t) \approx C(s,t) \quad \text{such that} \quad \frac{\partial C}{\partial t} \approx 0\,. \tag{45}$$

To comprehensively model the dispersion dynamics, this linear approximation must be modified by "nonlinear" terms arising from the varying curvature and torsion of the pipe and the axial gradients of the concentration. We account for such variations and gradients up to some specified order, and hence we derive a long-wave, slowly varying model for the dispersion in a long pipe.[6]

Thus the second stage is to seek iterative improvements to a given level of description of the center manifold and the low-dimensional evolution thereon. The aim is to find a low-dimensional description which satisfies the governing advection-diffusion equation (2). As in iterative methods for finding the zero of a function, we use the residual of the governing equation in order to guide corrections. The iteration scheme is successful as long as it ultimately drives the residual to zero to the desired order of accuracy; the center manifold approximation theorem [4] then assures us that the model is correct to the same asymptotic error. Suppose that at one stage of the iteration we have the approximate model (upper case Fraktur)

$$c \approx \mathfrak{C}(C) \quad \text{such that} \quad \frac{\partial C}{\partial t} \approx \mathfrak{G}(C)\,;$$

approximate because the residual of the governing differential equation (2)

$$R = \frac{\partial c}{\partial t} + \mathrm{Pe}\,\boldsymbol{v}\cdot\boldsymbol{\nabla} c - \nabla^2 c = \frac{\partial C}{\partial \mathfrak{C}}\mathfrak{G} - \mathcal{L}\mathfrak{C} - f(\mathfrak{C},\epsilon) = \mathcal{O}(\epsilon^q) \tag{46}$$

for some order of error q. We seek to find "small" corrections (indicated by lowercase Fraktur) so that

$$c \approx \mathfrak{C}(C) + \mathfrak{c}(C) \quad \text{such that} \quad \frac{\partial C}{\partial t} \approx \mathfrak{G}(C) + \mathfrak{g}(C)$$

is a better approximation to the center manifold and the evolution thereon by reducing the residual of the governing equation. The aim of each iteration is to improve the order of the error q so that, by the center manifold approximation theorem, we improve the accuracy of

[5] A quick argument is that the physical field and the evolution in any locale in the pipe is actually parametrized by the cross-sectional average concentration C and its axial gradients C_s, C_{ss}, etc., as *independent* "amplitudes" (as in parts of the calculus of variations). That C is differentiable implies we obtain the evolution of these independent amplitudes from $\partial C/\partial t = g(C)$: $\frac{\partial C_{s^n}}{\partial t} = \partial^n g/\partial s^n$. Then impose the simplest slowly varying paradigm that the nth derivative of C scales as δ^n for some small parameter δ corresponding to the typical axial gradient.

[6] A big advantage of the center manifold approach is that we avoid the daunting hierarchy of superslow space-time scales required by the method of multiple scales.

the model. Substituting into the governing differential equation (44) and using the chain rule for time derivatives lead to

$$\left(\frac{\partial\mathfrak{C}}{\partial C}+\frac{\partial\mathfrak{c}}{\partial C}\right)(\mathfrak{G}+\mathfrak{g})=\mathcal{L}\mathfrak{C}+\mathcal{L}\mathfrak{c}+f(\mathfrak{C}+\mathfrak{c},\epsilon)\,.$$

Given that it is impossible to solve this for the perfect corrections in one step, seek an approximate equation for the corrections of $\mathcal{O}(\epsilon^q)$ by

- ignoring products of corrections because they will be small, $\mathcal{O}(\epsilon^{2q})$, compared with the dominant effect of the linear correction terms;
- and replacing factors by their zeroth order approximation wherever they are multiplied by a correction, introducing errors $\mathcal{O}(\epsilon^{q+1})$, which slows the iteration convergence to linear, as opposed to the quadratic convergence which would be obtained otherwise (Geddes [15] gives an example of the quadratic convergence attained in the algebraic solution of simpler ODEs).

Thus we wish to solve

$$\frac{\partial\mathfrak{C}}{\partial C}\mathfrak{G}+\mathfrak{g}=\mathcal{L}\mathfrak{C}+\mathcal{L}\mathfrak{c}+f(\mathfrak{C},\epsilon)\,.$$

Rearranging and recognizing that $\frac{\partial\mathfrak{C}}{\partial C}\mathfrak{G}=\frac{\partial\mathfrak{C}}{\partial t}$ by the chain rule, we solve

$$\mathcal{L}\mathfrak{c}-\mathfrak{g}=\frac{\partial\mathfrak{C}}{\partial t}-\mathcal{L}\mathfrak{C}-f(\mathfrak{C},\epsilon) \tag{47}$$

for the corrections (on the left-hand side). The great advantage of this approach is that the right-hand side is simply the residual of the governing equation (44), here the advection-diffusion equation (2), evaluated at the current approximation. Thus at any iteration we just deal with physically meaningful expressions; all the complicated expansions and rearrangements of asymptotic expansions, as needed by the method of multiple scales [18, section 3.5] or earlier methods to find the center manifold [7, section 5.4], are absent. Evaluating the residual is very involved and potentially has enormous algebraic detail, much of which is repeated at every iteration. However, with the advent of computer algebra, as described in the next section, all this detail may be left to the computer to perform, whereas all a human need be concerned with is setting up the solution of (47) and not at all with the detailed algebraic machinations of asymptotic expansions.

The main detail is to solve equations of the form

$$\mathcal{L}\mathfrak{c}-\mathfrak{g}=R \tag{48}$$

for some given residual R. Recognize that its solution is not unique. The freedom comes from the fact that we may parametrize the center manifold in an almost arbitrary manner. The freedom is resolved by giving a precise meaning to the amplitudes; here we have set the amplitude C to be precisely the cross-section average concentration. To solve (48) we find it convenient to adopt the following procedure which is also familiar as part of other asymptotic methods. Rewrite (48) as $\mathcal{L}\mathfrak{c}=\mathfrak{g}+R$ and recognize that $\mathcal{L}$ is singular due to the zero eigenvalue of conservation of material in cross-pipe diffusion. We choose $\mathfrak{g}$ to place the right-hand side in the range of $\mathcal{L}$ by integrating over a cross-section. This is known as the solvability condition. Having put the right-hand side in the range of $\mathcal{L}$, we solve $\mathcal{L}\mathfrak{c} = \hat{R} = \mathfrak{g} + R$ for $\mathfrak{c}$, making the solution unique by requiring the cross-section averaged concentration to remain C. Then the last step of each iteration is to update the approximations for the center manifold shape and the evolution thereon.

When the residual is reduced to some asymptotic order, $R = \mathcal{O}(\epsilon^q)$, then the center manifold approximation theorem [4] assures us that the model is correct to the same order of error. Hence we write errors such as appear in (29). Due to the exponential quick decay to the center manifold, the relevance theorem [4] then indicates that the long-wave slowly varying dispersion models (4) and (35) apply after a few cross-pipe diffusion times.

Appendix B. Computer algebra derivation.

Computer algebra is a very powerful means to derive asymptotic expansions. In order for other people to reproduce and verify the results recorded herein, I list here the core of the program used to derive the asymptotic expansions; obtain the full program by request.

The computer algebra program was written in REDUCE[7] to calculate the asymptotic expansions of the center manifold models described in this article.

There is a lot of detail to the computer algebra program. However, the key to the correctness of the results is the coding of the governing equations which forms the key part of the core printed here. The algorithm iteratively drives the residuals of these equations to be smaller than some asymptotic orders of error; see Roberts [31] for a generic description of the algorithm. Thus the details about how the residuals are reduced are not vital; it is important only that they are correctly computed and are ultimately asymptotically negligible.

```
comment Find the flow in an arbitrarily curving pipe,
Simultaneously determine the shear dispersion in such a flow:
eps=magnitude of curvature terms,
del=magnitude of axial derivatives and of torsion.
;
depend kap,s; % curvature
depend tau,s; % torsion
% local coordinate system is (s,r,th), tp=th+phi(s)
depend tp,s,th;
let { df(tp,s)=>-tau, df(tp,th)=>1 };
hs:=1-eps*kap*r*cos(tp);
hr:=1;
ht:=r;
% trigonometry rules OK
let { sin(~a)*cos(~b) => (sin(a+b)+sin(a-b))/2
    , cos(~a)*cos(~b) => (cos(a-b)+cos(a+b))/2
    , sin(~a)*sin(~b) => (cos(a-b)-cos(a+b))/2
    , cos(~a)^2       => (1+cos(2*a))/2
    , sin(~a)^2       => (1-cos(2*a))/2
    };
% mean over a cross section (mult by r to use)
depend r,rt;depend tp,rt;
```

[7] At the time of writing, information about REDUCE was available from Anthony C. Hearn, RAND, Santa Monica, CA 90407-2138 (mailto:reduce@rand.org). There were demonstration versions of REDUCE freely available at ftp://ftp.zib.de/pub/reduce/demo or http://www.codemist.co.uk.

```
operator mean; linear mean;
let { mean(r^~m*cos(~n),rt) => 0
    , mean(r^~m*sin(~n),rt) => 0
    , mean(r^~m,rt) => 2/(m+1)
    , mean(r,rt) => 1
    };
% operators to solve for updates
...
% initial approximations
u:=2*(1-r^2) +eps*3/2*kap*(r-r^3)*cos(tp);
v:=0;
w:=0;
% pressure = ps + p
% = (local mean gradient) + (zero mean fluctuation)
ps:=-8; p:=0;
% concentration of tracer, mean c.
depend c,s,t;
let df(c,t)=>g;
cc:=c;
g:=0;
pe:=re*sc; % Peclet number = Reynolds * Schmidt
rh:=1; % approx reciprocal of axial scale factor hs

% iterate until residuals are negligible
let { eps^3=>0, del^5=>0 }; % truncate the asymptotics
repeat begin
    % reciprocal of scale factor
    eqr:=hs*rh-1;
    rh:=rh-eqr;
    % vorticity
    oms:=(df(r*w,r)-df(v,th))/r;
    omr:=(df(hs*u,th)-del*df(r*w,s))*rh/r;
    omt:=(del*df(v,s)-df(hs*u,r))*rh;
    % Navier-Stokes equation
    nss:=rh*(ps+del*df(p,s)) +(df(r*omt,r)-df(omr,th))/r
        +re*( del*u*df(u,s)*rh+v*df(u,r)+w*df(u,th)/r
             +u*df(hs,r)*v*rh+u*df(hs,th)*w*rh/r );
    nsr:=df(p,r)    +(df(hs*oms,th)-del*df(r*omt,s))*rh/r
        +re*( del*u*df(v,s)*rh+v*df(v,r)+w*df(v,th)/r
             -w^2/r-df(hs,r)*u^2*rh );
    nst:=df(p,th)/r +(del*df(omr,s)-df(hs*oms,r))*rh
        +re*( del*u*df(w,s)*rh+v*df(w,r)+w*df(w,th)/r
             -u^2*df(hs,th)*rh/r+v*w/r );
    % continuity equation
    cty:=(del*df(r*u,s)+df(r*hs*v,r)+df(hs*w,th))*rh/r;
    ...
    % equation for tracer evolution
    ceq:= df(cc,t) +pe*( del*u*df(cc,s)*rh+v*df(cc,r)+w*df(cc,th)/r )
         -(del^2*df(rh*r*df(cc,s),s)+df(r*hs*df(cc,r),r)
            +df(hs/r*df(cc,th),th))*rh/r;
    cmean:=mean(r*hs*cc,rt)-c;
    ...
end until (eqr=0)and(nss=0)and(nsr=0)and(nst=0)
       and(cty=0)and(ceq=0)and(cmean=0);
```

```

% check subsiduary conditions
uwall:=sub(r=1,u);
vwall:=sub(r=1,v);
wwall:=sub(r=1,w);
umean:=mean(r*u,rt);
pmean:=mean(r*p,rt);
cwall:=sub(r=1,df(cc,r));
cflux:=mean(r*(u*pe*cc-del*rh*df(cc,s)),rt);
end;
```

Observe that the pressure is decomposed into a mean gradient and a cross-pipe fluctuating component. There is code to adjust the mean pressure gradient to ensure a constant mean fluid flux. To adapt to the traditional fixed pressure gradient in a helical or toroidal pipe, one just needs to omit the adjustment. However, the results are then inappropriate to a pipe with varying curvature or torsion.

Acknowledgment. I thank the referees for their useful comments.

REFERENCES

[1] G. K. Batchelor, *An Introduction to Fluid Dynamics*, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.

[2] T. Beer and P. C. Toung, *Longitudinal dispersion in natural streams*, J. Environmental Engrg., 109 (1983), pp. 1049–1067.

[3] S. A. Berger, L. Talbot, and L.-S. Yao, *Flow in curved pipes*, Annu. Rev. Fluid Mech., 15 (1983), pp. 461–512.

[4] J. Carr, *Applications of Centre Manifold Theory*, Applied Math. Sci. 35, Springer-Verlag, New York, 1981.

[5] P. C. Chatwin, *The approach to normality of the concentration distribution of a solute in a solvent flowing along a straight pipe*, J. Fluid Mech., 43 (1970), pp. 321–352.

[6] P. C. Chatwin and C. M. Allen, *Mathematical models of dispersion in rivers and estuaries*, Annu. Rev. Fluid Mech., 17 (1985), pp. 119–149.

[7] P. H. Coullet and E. A. Spiegel, *Amplitude equations for systems with competing instabilities*, SIAM J. Appl. Math., 43 (1983), pp. 776–821.

[8] P. Daskopoulos and A. M. Lenhoff, *Flow in curved ducts: Bifurcation structure for stationary ducts*, J. Fluid Mech., 203 (1989), pp. 125–148.

[9] W. R. Dean, *Note on the motion of fluid in a curved pipe*, Phil. Mag., 4 (1927), pp. 208–223.

[10] W. R. Dean, *The stream-line motion of fluid in a curved pipe*, Phil. Mag., 5 (1928), pp. 673–695.

[11] M. E. Erdogan and P. C. Chatwin, *The effects of curvature and buoyancy on the laminar dispersion of solute in a horizontal tube*, J. Fluid Mech., 29 (1967), pp. 465–484.

[12] D. Fargie and B. W. Martin, *Developing flow in a pipe of circular cross section*, Proc. Roy. Soc. London A, 321 (1971), pp. 461–476.

[13] Th. Gallay, *A center-stable manifold theorem for differential equations in Banach spaces*, Comm. Math. Phys., 152 (1993), pp. 249–268.

[14] D. Gammack and P. E. Hydon, *Flow in pipes with non-uniform curvature and torsion*, J. Fluid Mech., 433 (2001), pp. 357–382.

[15] K. O. Geddes, *Convergence behaviour of the Newtonian iteration for first-order differential equations*, in Symbolic and Algebraic Computation, Lecture Notes in Comput. Sci. 72, E. W. Ng, ed., Springer-Verlag, New York, 1979, pp. 189–199.

[16] M. Germano, *On the effect of torsion on a helical pipe flow*, J. Fluid Mech., 125 (1982), pp. 1–8.

[17] S. Ghosal, *Lubrication theory for electro-osmotic flow in a microfluidic channel of slowly varying cross-section and wall charge*, J. Fluid Mech., 459 (2002), pp. 103–128.

[18] A. Jeffrey and T. Kawahara, *Asymptotic Methods in Nonlinear Wave Theory*, Applicable Mathematics Series, Pitman, Boston, 1982.

[19] M. Johnson and R. D. Kamm, *Numerical studies of steady flow dispersion at low Dean number in a gently curving tube*, J. Fluid Mech., 172 (1986), pp. 329–345.

[20] S. W. Jones and W. R. Young, *Shear dispersion and anomalous diffusion by chaotic advection*, J. Fluid Mech., 280 (1994), pp. 149–172.

[21] H. C. Kao, *Torsion effects on fully developed flow in a helical pipe*, J. Fluid Mech., 184 (1987), pp. 335–356.

[22] E. Kreyszig, *Advanced Engineering Mathematics*, 8th ed., Wiley, New York, 1999.

[23] J. Larrain and C. F. Bonilla, *Theoretical analysis of pressure drop in the laminar flow of fluid in a coiled pipe*, Trans. Soc. Rheol., 14 (1970), pp. 135–147.

[24] S. Liu and J. Masliyah, *Axially invariant laminar flow in helical pipes with finite pitch*, J. Fluid Mech., 251 (1993), pp. 315–353.

[25] D. J. McConalogue, *The effects of secondary flow on the laminar dispersion of an injected substance in a curved tube*, Proc. Roy. Soc. London A, 315 (1970), pp. 99–113.

[26] G. N. Mercer and A. J. Roberts, *A complete model of shear dispersion in pipes*, Japan J. Indust. Appl. Math., 11 (1994), pp. 499–521.

[27] S. Murata, Y. Miyake, and T. Inaba, *Laminar flow in a curved pipe with varying curvature*, J. Fluid Mech., 73 (1976), pp. 735–752.

[28] R. J. Nunge, T.-S. Lin, and W. N. Gill, *Laminar dispersion in curved tubes and channels*, J. Fluid Mech., 51 (1972), pp. 363–383.

[29] T. J. Pedley, *Fluid Mechanics of Large Blood Vessels*, Cambridge University Press, Cambridge, UK, 1980.

[30] A. J. Roberts, *The application of centre-manifold theory to the evolution of systems which vary slowly in space*, J. Austral. Math. Soc. Ser. B, 29 (1988), pp. 480–500.

[31] A. J. Roberts, *Low-dimensional modelling of dynamics via computer algebra*, Comput. Phys. Comm., 100 (1997), pp. 215–230.

[32] A. J. Roberts, *Low-dimensional modelling of dynamical systems applied to some dissipative fluid mechanics*, in Nonlinear Dynamics from Lasers to Butterflies, Lecture Notes in Complex Systems 1, R. Ball and N. Akhmediev, eds., World Scientific, Singapore, 2003, pp. 257–313.

[33] D. M. Ruthven, *The residence time distribution for ideal laminar flow in helical tubes*, Chem. Engrg. Sci., 26 (1971), pp. 1113–1121.

[34] R. Smith, *Longitudinal dispersion coefficients for varying channels*, J. Fluid Mech., 130 (1983), pp. 299–314.

[35] G. I. Taylor, *Dispersion of soluble matter in solvent flowing slowly through a tube*, Proc. Roy. Soc. London A, 219 (1953), pp. 186–203.

[36] G. I. Taylor, *Conditions under which dispersion of a solute in a stream of solvent can be used to measure molecular diffusion*, Proc. Roy. Soc. London A, 225 (1954), pp. 473–477.

[37] H. C. Topakoglu, *Steady laminar flows of an incompressible viscous fluid in curved pipes*, J. Math. Mech., 16 (1967), pp. 1321–1338.

[38] R. N. Trivedi and K. Vasudeva, *Axial dispersion in laminar flow in helical coils*, Chem. Engrg. Sci., 30 (1975), pp. 317–325.

[39] E. R. Tuttle, *Laminar flow in twisted pipes*, J. Fluid Mech., 219 (1990), pp. 545–570.

[40] K. Yamamoto, S. Yanase, and T. Yoshida, *Torsion effect on the flow in a helical pipe*, Fluid Dynam. Res., 14 (1994), pp. 259–273.

[41] K. Yamamoto, S. Yanase, and T. Yoshida, *Erratum: Torsion effect on the flow in a helical pipe*, Fluid Dynam. Res., 24 (1999), pp. 309–311.

[42] L. Zabielski and A. J. Mestel, *Steady flow in a helically symmetric pipe*, J. Fluid Mech., 370 (1998), pp. 297–320.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 463–502

Near-Resonant Steady Mode Interaction: Periodic, Quasi-periodic, and Localized Patterns[*]

María Higuera[†], Hermann Riecke[‡], and Mary Silber[‡]

Abstract. Motivated by the rich variety of complex periodic and quasi-periodic patterns found in systems such as two-frequency forced Faraday waves, we study the interaction of two spatially periodic modes that are *nearly resonant*. Within the framework of two coupled one-dimensional Ginzburg–Landau equations we investigate analytically the stability of the periodic solutions to general perturbations, including perturbations that do not respect the periodicity of the pattern, and which may lead to quasi-periodic solutions. We study the impact of the deviation from exact resonance on the destabilizing modes and on the final states. In regimes in which the mode interaction leads to the existence of traveling waves our numerical simulations reveal localized waves in which the wavenumbers are resonant and which drift through a steady background pattern that has an off-resonant wavenumber ratio.

Key words. pattern formation, mode interaction, quasi-periodic, parity breaking, longwave instability, shortwave instability, localized drift waves

AMS subject classifications. 37L05, 35B10, 35B15, 35B34, 35B35

DOI. 10.1137/030600552

1. Introduction.

Pattern-forming instabilities lead to an astonishing variety of spatial and spatio-temporal structures, ranging from simple periodic stripes (rolls) to spatially localized structures and spatio-temporally chaotic patterns. Even within the restricted class of steady spatially ordered patterns a wide range of patterns has been identified and investigated beyond simple square or hexagonal planforms including patterns exhibiting multiple length scales: superlattice patterns, in which the length scales involved are rationally related rendering the pattern periodic albeit on an unexpectedly large length scale, and quasi patterns, which are characterized by incommensurate length scales and which are therefore not periodic in space. These more complex two-dimensional patterns have been observed in particular in the form of Faraday waves on a fluid layer that is vertically shaken with a two-frequency periodic acceleration function [17, 1, 26, 2] and to some extent also in vertically vibrated Rayleigh–Bénard convection [44] and in nonlinear optical systems [35]. The Faraday system is especially suitable for experimental investigation of pattern formation in systems with two competing

[*] Received by the editors July 26, 2003; accepted for publication (in revised form) by B. Fiedler April 26, 2004; published electronically November 9, 2004. The work of María Higuera was supported by the Spanish Ministerio de Ciencia y Tecnologí́a (BFM2001-2363). The work of Hermann Riecke was supported by NASA grant NAG3-2113, the Department of Energy (DE-FG02-92ER14303) and NSF (DMS-9804673). The work of Mary Silber was supported by NASA grant NAG3-2364, and by NSF under DMS-9972059 and the MRSEC Program under DMR-0213745.

http://www.siam.org/journals/siads/3-4/60055.html

[†] E. T. S. Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, Plaza Cardenal Cisneros 3, 28040 Madrid, Spain (maria@fmetsia.upm.es).

[‡] Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208 (h-riecke@northwestern.edu, m-silber@northwestern.edu).

spatial modes of instability since such codimension-two points are easily accessible by simply adjusting the frequency content of the periodic forcing function [17, 5]. The observation of both superlattices and quasi patterns in this physical system raises an intriguing question concerning the selection of these kinds of patterns: What determines whether, for given physical parameters, a periodic or a quasi-periodic pattern is obtained? This provides the main motivation for the present paper in which we address certain aspects of the selection problem within the somewhat simple framework of mode competition in one spatial dimension.

To capture the competition between commensurate/incommensurate length scales we focus on the interaction between two modes with a wavelength ratio that is close to, but not necessarily equal to, the ratio of two small integers. We are thus led to consider a *near-resonant* mode interaction. We focus on the case in which both modes arise in a steady bifurcation. While at first sight it may seem that an analysis of steady-state modes would not be applicable to two-frequency forced Faraday waves, for which most of the patterns of interest in the present context have been observed, it should be noted that very close to onset the amplitude equations for these waves can be reduced to the standing-wave subspace, in which the waves satisfy equations that have the same form as those for modes arising from a steady bifurcation (see, for example, [38]).

The competition between commensurate and incommensurate steady structures has been addressed previously in the context of a spatially periodic forcing of patterns [29, 7, 10, 12, 34]. Using a one-dimensional *external* forcing of two-dimensional patterns in electroconvection of nematic liquid crystals, localized domain walls in the local phase of the patterns were observed if the forcing wavenumber was sufficiently incommensurate with the preferred wavenumber of the pattern [29]. Theoretically, the domain walls were described using a one-dimensional Ginzburg–Landau equation that included a near-resonant forcing term, which reflects the small mismatch between the forcing wavenumber and the wavenumber of the spontaneously forming pattern [7, 8]. The situation we have in mind in the present paper is similar to these studies in so far as the competing modes we investigate also provide a periodic forcing for each other. The case of external forcing is recovered if one of the two modes is much stronger than the other and consequently the feedback from the forced mode on the forcing mode can be ignored. We do not restrict ourselves to this case, however, and thus both modes are active degrees of freedom, and the interaction between them is mutual.

The analysis of the interaction of two exactly resonating modes near a codimension-two point at which both modes bifurcate off the basic, unpatterned state has revealed a wide variety of patterns and dynamical phenomena. Rich behavior has been found in small systems in which the interacting modes are determined by the symmetry and shape of the physical domain (e.g., [47, 31, 11, 20, 48, 33, 24]). More relevant for our goal are the studies of mode-interaction in the presence of translation symmetry, since they allow the extension to systems with large aspect ratio, which are required to address the difference between commensurate and incommensurate structures. Our investigation builds on a comprehensive analysis, performed by Dangelmayr [13], of the interaction between two resonant spatial modes in $O(2)$-symmetric systems with wavenumbers in the ratio $m : n$, $m < n$. (Here the $O(2)$-symmetry is a consequence of restricting our attention to spatially periodic patterns in a translation-invariant system.) For $m > 1$ there are two primary bifurcations off the trivial state to pure modes followed by secondary bifurcations to mixed modes. These mixed modes can in turn undergo a Hopf bifurcation to generate standing wave solutions and a parity-breaking bifurcation that produces traveling waves. For $m = 1$ similar results are found except that there is only one pure-mode state—the one with the higher wavenumber; the other primary bifurcation leads directly to a branch of mixed modes. Further analyses of this system have revealed structurally and asymptotically stable heteroclinic cycles near the mode interaction point when $m : n = 1 : 2$ [25, 39, 3]. More recently, complex dynamics organized around a sequence of transitions between distinct heteroclinic cycles has been discovered in resonances of the form $1 : n$ [36, 37], in particular, in the cases $n = 2$ and $n = 3$.

Although early work on resonant mode interactions considered only strictly periodic solutions, it provided insight into various phenomena that were observed experimentally in large-aspect ratio systems involving large-scale modulations of the patterns. For example, in steady Taylor vortex flow it was found experimentally that not too far from threshold the band of experimentally accessible wavenumbers is substantially reduced compared to the stability limit obtained by the standard analysis of side-band instabilities in the weakly nonlinear regime [15]. The origin of this strong deviation was identified to be a saddle-node bifurcation associated with the $1 : 2$ mode interaction [42]. In directional solidification [46] localized drift waves have been observed, which arise from the parity-breaking bifurcation [30, 9, 22] that is associated with the resonant mode interaction with wavenumber ratio $1 : 2$ [28]. Subsequently such waves have also been obtained in a variety of other systems including directional viscous fingering [40], Taylor vortex flow [49, 43], and premixed flames [4]. In our treatment of the near-resonant case the mode amplitudes are allowed to vary slowly in space. It therefore naturally incorporates phenomena like the localized drift waves and the modification of side-band instabilities by the resonance. The interaction of two one-dimensional steady patterns with different natural spatial scales in large reflection symmetric domains has been studied in [32, 16]. Specifically, in [32] an analysis is carried out on a nonlocal system of equations describing the interaction between the modulation of a shortwave mode and a longwave mode. The authors show that in addition to mixed-mode solutions traveling waves can exist in a large region of parameter space. In [16], a pair of singularly perturbed Ginzburg–Landau equations, truncated at cubic order, is considered to describe the interaction that occurs between two shortwave modes with modulation scales differing significantly. These equations are shown to contain a very complicated set of localized stationary patterns.

In this paper we study the interaction of two nearly resonant modes in a spatially extended, driven, dissipative system. Near onset we model the slow dynamics of such systems by two amplitude equations of Ginzburg–Landau type, one for each mode. We focus on the weak resonances ($m + n \geq 5$) in order to avoid some of the specific features of the strong-resonance cases (for example, the structurally stable heteroclinic cycles in the case $m : n = 1 : 2$). A recent paper by Dawes, Postlethwaite, and Proctor presents a complementary investigation near the $1 : 2$ resonance [14]. Our primary goal here is to investigate the transitions between periodic and quasi-periodic states that take place as the result of side-band instabilities, with an eye on how the detuning from exact spatial resonance influences this process. We find that the detuning can play an important role in the selection of the final wavenumbers of the modes involved. For example, it can favor a periodic to quasi-periodic transition that would otherwise (i.e., in the case of exact resonance) result in a second periodic state. Among the various quasi-periodic states that we find in numerical simulations are several that consist of drifting localized structures with alternating locked (periodic) and unlocked (quasi-periodic) domains.

It should be noted that at present there is no rigorous justification for the description of quasi-periodic patterns using low-order amplitude expansions. In fact, the lack of straightforward convergence of such an expansion has been investigated recently for two-dimensional quasi patterns [45]. We will not discuss these issues; instead we use the coupled Ginzburg–Landau equations as model equations that are known to be the appropriate equations for periodic patterns and at the same time also allow quasi-periodic solutions.

The organization of the paper is as follows. In section 2 we set up the coupled Ginzburg–Landau equations that are based on the truncated normal form equations for the $m : n$ resonance. In section 3 we utilize and build upon the detailed results of Dangelmayr [13] to describe the stability properties of steady spatially periodic states (i.e., pure and mixed modes) with respect to perturbations that preserve the periodicity of the pattern. In section 4 we turn to the question of stability of the steady periodic solutions with respect to side-band instabilities. Here we also determine how these instabilities are affected by the detuning from perfect resonance. In section 5 we carry out numerical simulations to investigate the nonlinear evolution of the system subsequent to the side-band instability identified in section 4. The numerical investigations presented in sections 4 and 5 focus on the case $m : n = 2 : 5$. This case is representative of weak resonances. Moreover, it possesses an additional interesting feature in that the resonance terms (i.e., the terms that couple the phases of the two modes) can affect the stability of primary solutions. Finally, our concluding remarks are given in section 6.

2. The amplitude equations.

We consider driven dissipative systems in one spatial dimension that are invariant under spatial translations and reflections. We further assume that in the system of interest there are two distinct spatial modes that destabilize the basic homogeneous state nearly simultaneously. The wavenumbers of these two modes, q_1 and q_2, correspond to minima of the neutral curves and are assumed to be in approximate spatial resonance, i.e.,

$$n(q_1 + \varepsilon\hat{\gamma}) = mq_2, \qquad |\varepsilon| \ll 1, \tag{2.1}$$

where m and $n > m$ are positive coprime integers and the term $\varepsilon\hat{\gamma}$ measures the deviation from perfect resonance. Physical fields, near onset, may then be expanded in terms of these two spatial modes:

$$u(x,t) = \varepsilon[A_1(X,T)e^{i(q_1+\varepsilon\hat{\gamma})x} + A_2(X,T)e^{iq_2x} + \text{c.c.}] + \cdots . \tag{2.2}$$

The two modes are allowed to vary on slow spatial and temporal scales $X = \varepsilon x$ and $T = \varepsilon^2 t$, respectively. Note that in the expansion (2.2) we do not expand about the minima $q_{1,2}$ of the neutral stability curves, which for $\hat{\gamma} \neq 0$ are not in spatial resonance, but take instead a mode A_1 which *is* in exact spatial resonance with A_2. This choice of A_1 and A_2 simplifies the equations, avoiding any explicit dependence on the spatial variable in the resulting amplitude equations. Furthermore, we allow for a small offset between the critical values of the forcing amplitudes F_{1c}, F_{2c} of the two modes (see Figure 1).

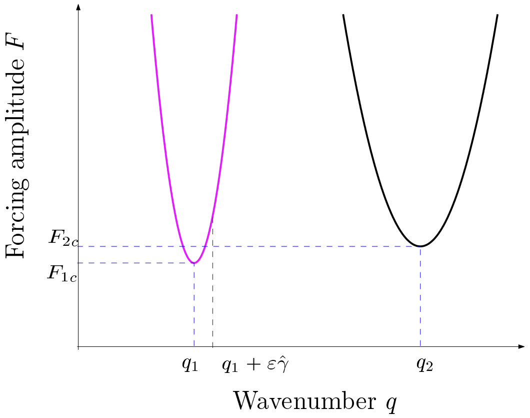


Figure 1. *Sketch of the neutral stability curve for the two modes.*

The equations governing the evolution of A_1 and A_2 must be equivariant under the symmetry operations generated by spatial translations (T_φ) and spatial reflections (R), which act on (A_1, A_2) as follows:

$$\begin{aligned} T_\varphi &: (A_1, A_2) \to (e^{im\varphi}A_1, e^{in\varphi}A_2) \quad \text{for} \quad \varphi \in [0, 2\pi), \\ R &: (A_1, A_2) \to (\bar{A}_1, \bar{A}_2), \end{aligned} \tag{2.3}$$

where the bar denotes the complex conjugate. Consistent with this equivariance requirement, the slow evolution of A_1 and A_2 can be approximated, after rescaling, by the Ginzburg–Landau equations

$$A_{1T} = \mu A_1 + \delta A_{1XX} - i\gamma A_{1X} - (s|A_1|^2 + \rho|A_2|^2)A_1 + \nu \bar{A}_1^{n-1} A_2^m, \tag{2.4}$$

$$A_{2T} = (\mu + \Delta\mu)A_2 + \delta' A_{2XX} - (s'|A_2|^2 + \rho'|A_1|^2)A_2 + \nu' \bar{A}_2^{m-1} A_1^n. \tag{2.5}$$

The subscripts indicate partial derivatives with respect to X and T. The main control parameter $\mu \propto (F - F_{1c})/\varepsilon^2$ measures the magnitude of the overall forcing. In addition, we keep track of the offset in the two critical forcing amplitudes with $\Delta\mu \equiv (F_{1c} - F_{2c})/\varepsilon^2$ (see Figure 1) and capture the detuning between q_1 and mq_2/n with $\gamma \equiv 2\delta\hat{\gamma}$. The local curvature of the neutral stability curves near q_1 and q_2 is measured by δ and δ', respectively. We further assume that the nonlinear self- and cross-interaction coefficients satisfy the nondegeneracy conditions $ss' \neq 0$ and $ss' - \rho\rho' \neq 0$ and perform a simple rescaling such that $s = \pm 1$, $s' = \pm 1$.

One goal of this paper is to gain insight into the difference between periodic and quasi-periodic patterns in systems with two unstable wavenumbers. If the two wavenumbers are not rationally related and their irrational ratio is kept fixed as the onset for the two modes is approached, then only terms of the form $A_i|A_j|^{2l}$, $l = 1, 2, 3 \ldots$, appear in the equation for A_i. For a rational ratio, however, additional nonlinear terms arise, which couple the otherwise uncoupled phases of the two modes A_i. For the $m : n$-resonance the leading-order resonance terms are given by $\bar{A}_1^{n-1}A_2^m$ and $\bar{A}_2^{m-1}A_1^n$ in the equations for A_1 and A_2, respectively. In order to explore the connection between the rational and the irrational case, we consider a wavenumber ratio which may be irrational, but its deviation from the ratio $m : n$ is of $\mathcal{O}(\varepsilon)$. This allows the mismatch between the two wavenumbers to be captured by the slow spatial variable X and allows us to describe periodic and quasi-periodic patterns with the same set of equations (2.4), (2.5). Equivalently, we could have expanded in the irrationally related wavenumbers q_1 and q_2 associated with the minima of the neutral curves in Figure 1. Then the resonance terms would introduce space-periodic coefficients with a period that is related to the mismatch of the wavenumbers. Our choice of the expansion wavenumbers (cf. (2.1)) removes this space-dependence and introduces the first-order derivative $-i\gamma\partial_X A_1$ in its place.

We focus on the weak resonances, $m + n \geq 5$, in which the resonant terms are of higher order. We neglect, however, nonresonant terms of the form $A_i|A_j|^p$ ($i, j = 1$, 2 and $4 \leq p \in \mathbb{N}$) which may arise at lower order. This is motivated by the observation that such terms do not contribute any qualitatively new effects for small amplitudes. The resonant terms, in contrast, remove the unphysical degeneracy that arises when the phases are left uncoupled and can therefore influence dynamics in a significant way despite appearing at higher order. Note, however, that near onset the resonant terms are typically small and the phase coupling between A_1 and A_2 occurs on a very slow time scale. The coupling becomes stronger further above onset where the weakly nonlinear analysis may no longer be valid.

It is often useful to recast (2.4), (2.5) in terms of real amplitudes $R_j \geq 0$ and phases ϕ_j by writing $A_j = R_j e^{i\phi_j}$. This leads to

$$\begin{aligned} R_{1T} &= \mu R_1 - (sR_1^2 + \rho R_2^2)R_1 + \nu R_1^{n-1}R_2^m \cos(n\phi_1 - m\phi_2) \\ &\qquad + \delta R_{1XX} - \delta\phi_{1X}^2 R_1 + \gamma\phi_{1X}R_1, \end{aligned} \tag{2.6}$$

$$\begin{aligned} R_{2T} &= (\mu + \Delta\mu)R_2 - (s'R_2^2 + \rho' R_1^2)R_2 + \nu' R_2^{m-1}R_1^n \cos(n\phi_1 - m\phi_2) \\ &\qquad + \delta' R_{2XX} - \delta'\phi_{2X}^2 R_2, \end{aligned} \tag{2.7}$$

$$R_1\phi_{1T} = -\nu R_1^{n-1}R_2^m \sin(n\phi_1 - m\phi_2) + \delta\phi_{1XX}R_1 + 2\delta\phi_{1X}R_{1X} - \gamma R_{1X}, \tag{2.8}$$

$$R_2\phi_{2T} = \nu' R_2^{m-1}R_1^n \sin(n\phi_1 - m\phi_2) + \delta'\phi_{2XX}R_2 + 2\delta'\phi_{2X}R_{2X}. \tag{2.9}$$

If the spatial dependence in (2.6)–(2.9) is ignored, the system reduces to a set of ODEs equivalent to the one analyzed by Dangelmayr [13]. In this simplified problem, the translational symmetry (T_φ) causes the overall phase to decouple and leaves only the three real variables: R_1, R_2, and the mixed phase

$$\phi = n\phi_1 - m\phi_2, \tag{2.10}$$

with dynamically important roles. Dangelmayr's bifurcation analysis produced expressions for the location of primary bifurcations to pure-mode solutions, secondary bifurcations to mixed-mode solutions, and, in some instances, tertiary bifurcations to standing-wave and traveling-wave solutions. These results apply to a general $m : n$ resonance and prove useful in what follows.

3. Steady spatially periodic solutions.

In this section we analyze steady solutions of (2.4), (2.5) of the form

$$A_1 = R_1 e^{i(kX + \hat{\phi}_1)}, \quad A_2 = R_2 e^{i((nk/m)X + \hat{\phi}_2)}, \tag{3.1}$$

where $R_{1,2} \geq 0$ and $\hat{\phi}_{1,2}$ and k are real. Such states represent spatially periodic solutions of the original problem with wavenumbers $\tilde{q}_1 = q_1 + \varepsilon\gamma + \varepsilon k$ and $\tilde{q}_2 = q_2 + \varepsilon nk/m$ so that $\tilde{q}_1 n = \tilde{q}_2 m$. These solutions break the continuous translational symmetry (T_φ) but remain invariant under discrete translations.

Within this family of steady states there are generically only two types of nontrivial solutions, pure modes and mixed modes, which we describe below.

I. Pure modes ($S_{1,2}$). These are single-mode states, which take one of two forms:

$$\begin{aligned} &S_1 : (A_1, A_2) = (\sqrt{\alpha/s}\ e^{i(kX+\hat{\phi}_1)}, 0) \quad \text{for } m > 1, \\ &S_2 : (A_1, A_2) = (0, \sqrt{\beta/s'}\ e^{i((nk/m)X+\hat{\phi}_2)}), \end{aligned} \tag{3.2}$$

where $\hat{\phi}_1, \hat{\phi}_2 \in [0, 2\pi)$ and

$$\begin{aligned} \alpha &= \mu - \delta k^2 + \gamma k, \\ \beta &= \mu + \Delta\mu - \delta'(nk/m)^2. \end{aligned} \tag{3.3}$$

Note that pure modes of type S_1 are not present if $m = 1$ (see (2.4), (2.5)). Moreover, the pure modes S_1 and S_2 are not isolated but emerge as circles of equivalent solutions (parametrized by $\hat{\phi}_1$ or $\hat{\phi}_2$). Hereafter we consider resonances $m : n$, where $m \geq 2$, in which case both pure modes S_1 and S_2 are present.

II. Mixed modes ($S_\pm$). There are two types of mixed modes,

$$S_\pm : (A_1, A_2) = (R_1\ e^{i(kX+\hat{\phi}_1)}, R_2\ e^{i((nk/m)X+\hat{\phi}_2)}), \tag{3.4}$$

satisfying

$$\begin{aligned} S_\pm : \quad &\cos(\phi) = \pm 1, \\ &(s'\alpha - \rho\beta) = (ss' - \rho\rho')R_1^2 \pm (s'\nu R_2^2 - \nu'\rho R_1^2)R_2^{m-2}R_1^{n-2}, \\ &(s\beta - \rho'\alpha) = (ss' - \rho\rho')R_2^2 \pm (s\nu' R_1^2 - \nu\rho' R_2^2)R_2^{m-2}R_1^{n-2}. \end{aligned} \tag{3.5}$$

Here ϕ is the mixed phase given by (2.10), and α and β are defined by (3.3). As in the case of the pure modes $S_{1,2}$, translational symmetry implies that there are circles of equivalent mixed-mode states (parametrized by $\hat{\phi}_1$, say). Like the pure modes, the mixed modes are invariant under reflections (R) through an appropriate origin.

3.1. Stability under homogeneous perturbations. The stability of $S_{1,2}$ and $S_\pm$ under homogeneous perturbations can be obtained from Dangelmayr's analysis [13], which we review here in some detail to provide the background necessary for our analysis. The stability regions in the (α, β)-unfolding plane simply need to be mapped to the (k, μ)-plane with the (nonlinear) transformation (3.3). Each intersection of the curves $\alpha = 0$ and $\beta = 0$ corresponds to the codimension-two point of [13]; there are generically zero or two such intersections. Since the nonlinear coefficients are identical in the vicinity of both intersections, the response of the system to homogeneous perturbations in corresponding neighborhoods of the (two) intersections is identical. In particular, any bifurcation set arising from one intersection arises from the

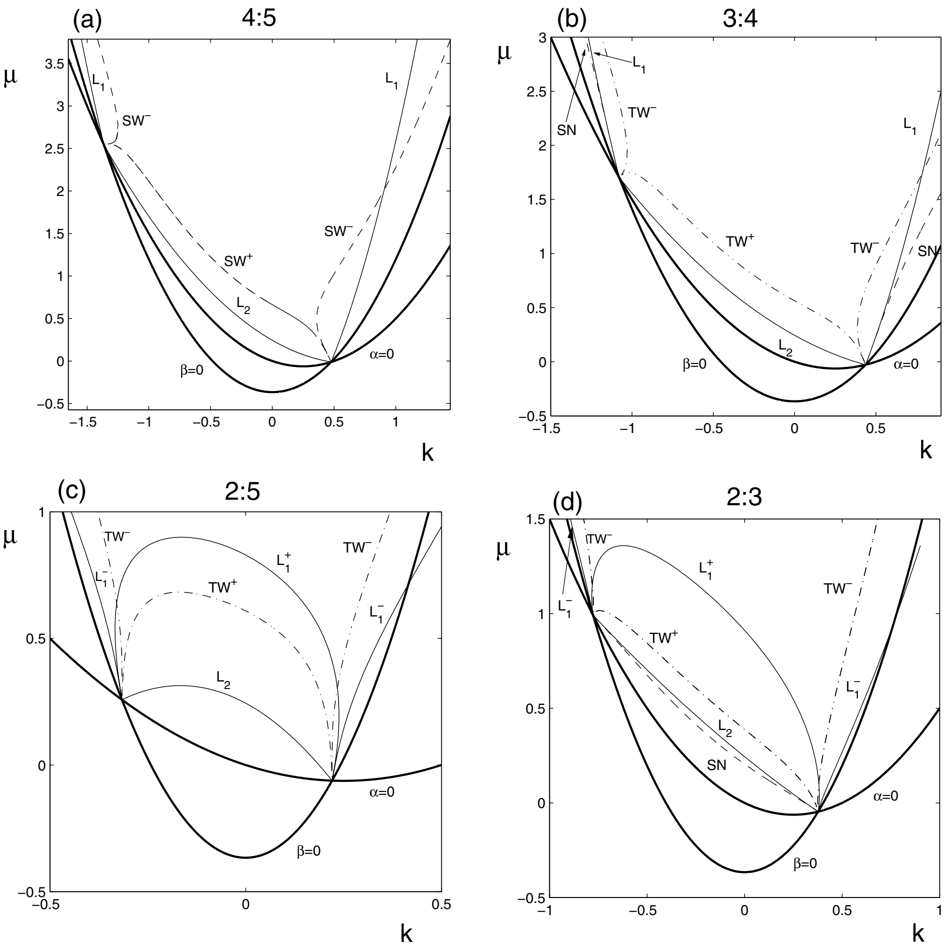


Figure 2. *Bifurcation sets indicating the creation of pure modes (thick solid lines), mixed modes (thin solid lines $L_1^{\pm}$, L_1, and L_2), standing waves (dashed lines $SW^{\pm}$), and traveling waves (dash-dotted lines $TW^{\pm}$), as well as saddle-node bifurcations (dashed lines SN). The specific resonance $m : n$ is indicated above the plots. In each case $\Delta\mu = 0.5$, $\gamma = 0.5$, $\delta = \delta' = 1$, $\rho = 0.4$, and $\rho' = 0.67$. In* (a) *$s = -s' = 1$, $\nu = 0.62$, and $\nu' = 1.02$, but in* (b), (c), *and* (d) *$s = s' = -1$, $\nu = 0.62$, and $\nu' = -1.02$.*

other as well. Several examples are given in Figure 2, which shows the various bifurcation sets in the (k, μ)-plane in four representative cases. These bifurcations are described below. Note that for mode A_1 the deviation of the wavenumber from the critical wavenumber is given by εk, whereas for mode A_2 it is given by $\varepsilon nk/m$. Additional details, valid in sufficiently small neighborhoods of the intersections, are available in [13].

In this paper we study (2.4), (2.5) with the coefficients ν and ν' of the higher-order resonance terms taken to be of order 1. The resulting stability and bifurcation diagrams therefore contain certain features that do not remain local to the bifurcation when $\nu, \nu' \to 0$; i.e., in this limit these features disappear at infinity, $\mu \to \infty$, and do not represent robust aspects of the mode-interaction problem. We return to this issue briefly at the end of this section and indicate which features of our sample bifurcation sets are not robust in the limit

$\nu, \nu' \to 0$.

The pure modes S_1 and S_2 bifurcate from the trivial state when $\alpha = 0$ and $\beta = 0$, respectively; the bifurcation to S_1 (S_2) is supercritical if $s = 1$ ($s' = 1$). Their stability is determined by four eigenvalues, one of which is forced to be zero by translation symmetry. In the case of S_1 the remaining three eigenvalues are

$$\lambda_0^{(1)} = -2\alpha/s,$$
$$L_1 : \ \lambda_\pm^{(1)} = (\beta - \rho'\alpha)/s \pm \left\{\nu' \, |\alpha/s|^{n/2}\right\}_{m=2}, \tag{3.6}$$

where the bracketed term with subscript $m = 2$ is present only if $m = 2$. For S_2 the eigenvalues are given by

$$\lambda_0^{(2)} = \beta/s',$$
$$L_2 : \ \lambda_+^{(2)} = \lambda_-^{(2)} = (s'\alpha - \rho\beta)/s'. \tag{3.7}$$

When one of the eigenvalues $\lambda_\pm^{(1)}$ ($\lambda_\pm^{(2)}$) changes sign the pure modes S_1 (S_2) become unstable to mixed modes, respectively. For S_2 the two eigenvalues coincide, and the bifurcation occurs along a single line denoted by L_2 in Figure 2. Similarly, for $m > 2$ the transition from S_1 to the mixed modes $S_\pm$ occurs along the single curve L_1 (Figure 2(a),(b)). For $m = 2$ the eigenvalues $\lambda_\pm^{(1)}$ are not degenerate and the line L_1 splits into two curves, L_1^+ and L_1^-; the mixed mode S_+ bifurcates at L_1^+ and S_- bifurcates at L_1^- ($L_1^\pm$ in Figures 2(c),(d)). The splitting is due to the presence of the resonant terms which are linear in A_2 if $m = 2$ but not otherwise. Both curves, L_1^+ and L_1^-, become tangent to each other at the intersection $\alpha = \beta = 0$.

The response of the mixed modes to amplitude perturbations is decoupled (due to reflection symmetry) from the effect of phase perturbations. The amplitude stability is determined by the eigenvalues of a 2 × 2-matrix $M_\pm$, whose determinant and trace can be written as (recall we consider $m > 1$)

$$\det(M_\pm) = -4(\rho\rho' - ss')R_1^2 R_2^2 \pm R_1^{n-2} R_2^{m-2} H(R_1, R_2), \tag{3.8}$$
$$\mathrm{Tr}(M_\pm) = -2s(R_1^2 + ss'R_2^2) \pm R_1^{n-2} R_2^{m-2} (\nu(n-2)R_2^2 + \nu'(m-2)R_1^2) \tag{3.9}$$

with

$$H(R_1, R_2) = -2\left[\nu' s(m-2)R_1^4 + \nu s'(n-2)R_2^4 - R_1^2 R_2^2(\rho'\nu m + \rho\nu' n)\right]. \tag{3.10}$$

Here R_1 and R_2 are solutions of (3.5). Since in general ν and ν' are of $\mathcal{O}(\varepsilon^{m+n-4})$, the contribution $H(R_1, R_2)$ from the resonance term affects the steady bifurcation determined by (3.8) only for $m \leq 3$. In particular, in the cases $m : n = 3 : n$ and $m : n = 2 : 3$ the function H can balance the first term along curves through the codimension-two point along which $R_2 \ll R_1$ and $R_1 \ll R_2$, respectively, and one of the mixed states experiences a saddle-node bifurcation (curves labeled SN in Figures 2(b),(d)) [13]. For $2 : n$ resonances with $n \geq 5$ the term R_1^4 drops out of $H(R_1, R_2)$. Consequently, the H-term cannot balance the first term in (3.8) and no saddle-node bifurcations occur.

For $m \geq 4$ the sign of $\det(M_\pm)$ does not depend on $R_{1,2}$. Then the $S_\pm$ solutions are always unstable for $(\rho\rho' - ss') > 0$, while for $(\rho\rho' - ss') < 0$ their stability must be deduced from the sign of $\mathrm{Tr}(M_\pm)$. If $ss' = 1$, it follows from (3.9) that for $\varepsilon \ll 1$ $\mathrm{sign}(\mathrm{Tr}(M_\pm)) = -s$; the mixed modes $S_\pm$ are then stable to amplitude perturbations if $s = +1$. On the other hand, if $ss' = -1$, the trace of $M_\pm$ can change sign, indicating the possibility of a Hopf bifurcation. The resulting time-periodic solutions inherit the reflection symmetry of $S_\pm$ (so $\dot{\phi}_1 = \dot{\phi}_2 = 0$) and therefore correspond to standing waves (SW curves in Figure 2(a)).

Instabilities associated with perturbations of the mixed phase (2.10) lead to bifurcations breaking the reflection symmetry. The relevant eigenvalues are given by

$$TW: \; e_\pm = \mp(\nu n R_2^2 + \nu' m R_1^2) R_1^{n-2} R_2^{m-2} \tag{3.11}$$

and may pass through zero only if $\mathrm{sign}(\nu\nu') = -1$. In this case the $S_\pm$ states undergo a pitchfork bifurcation, reflection symmetry is broken, and traveling waves appear (TW curves in Figures 2(b),(d)). These traveling-wave solutions manifest themselves as fixed points of the three-dimensional ODE system involving R_1, R_2, and ϕ but are seen to be traveling waves by the fact that the individual phase velocities are nonzero: $\dot{\phi}_1/\dot{\phi}_2 = n/m$. Since the phase velocity of these waves goes to 0 at the bifurcation, they are often called drift waves. We do not consider the stability properties of the traveling-wave solutions in this paper.

The stability results for $S_{1,2}$ and $S_\pm$ described above are illustrated in Figures 3, 4, and 5 for $m : n = 2 : 5$ and the indicated parameter values. They all satisfy

$$s = s' = 1, \qquad ss' - \rho\rho' > 0, \qquad \nu\nu' < 0 \tag{3.12}$$

so that the pure modes $S_{1,2}$ bifurcate supercritically in all cases. Next to these plots we sketch the type of bifurcation diagram one obtains when increasing μ at constant k along the thin dashed vertical lines. Since in all cases $ss' - \rho\rho' > 0$, both mixed modes $S_\pm$ are stable to amplitude perturbations (see the explanation following (3.8), (3.9)). The stability of $S_\pm$ with regard to phase perturbations depends, however, on the eigenvalue $e_\pm$ given by (3.11). Because $\nu\nu' < 0$, this eigenvalue may change sign, causing the mixed modes $S_\pm$ to undergo a symmetry-breaking bifurcation to traveling waves. Figures 3, 4, and 5 present six different cases characterized by the following quantities:

$$\chi = \gamma^2 - 4\Delta\mu\left(\delta - \left(\frac{n}{m}\right)^2 \delta'\right) \quad \text{and} \quad \Lambda = \frac{\gamma^2}{4\delta} - \Delta\mu. \tag{3.13}$$

The parameter χ controls the intersection of the parabolas $\alpha = 0$ and $\beta = 0$; they intersect if $\chi \geq 0$ and not otherwise. The quantity Λ determines the relative position (μ-value) of the minima of the curves $\alpha = 0$ and $\beta = 0$. It thus indicates which of the two modes, S_1 or S_2, is excited first. S_2 appears first when $\Lambda > 0$, while S_1 takes priority for $\Lambda < 0$; if $\Lambda = 0$, both pure modes onset simultaneously.

The degenerate case $\chi = \Lambda = 0$, i.e., $\gamma = \Delta\mu = 0$, is illustrated in Figure 3. In this case the curves $\alpha = 0$ and $\beta = 0$ intersect only once and their minima coincide. While in Figure 3(a) the neutral curve for mode A_1 is wider than that for A_2, it is the other way around in Figure 3(b). Note that the wavenumber nk/m of A_2 is larger than that of A_1; therefore to make the neutral curve of A_2 wider than that of A_1 requires a large ratio of δ/δ'. Depending

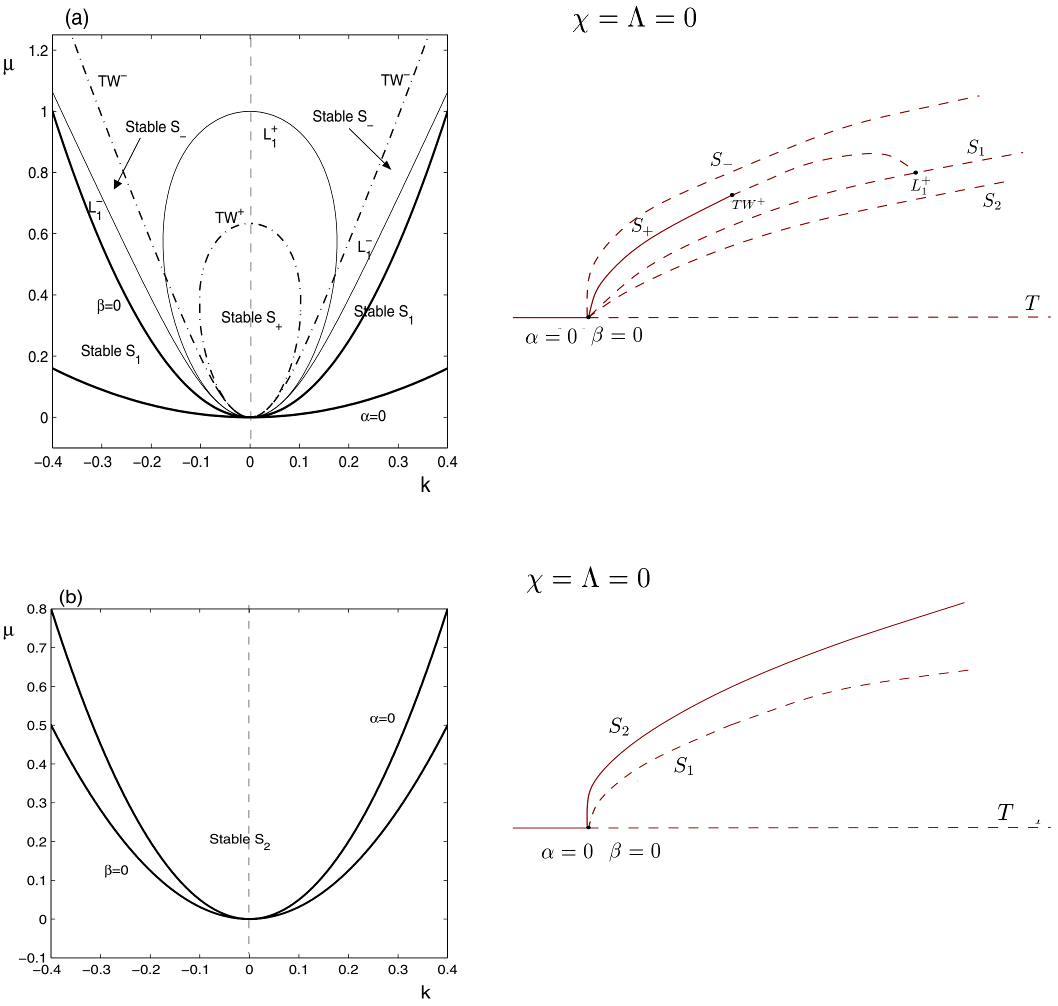


Figure 3. *Stability regions of the pure modes and mixed modes for the resonance* $2:5$ *for* $\Delta\mu = 0$ *and* $\gamma = 0$*. Figures on the right are sketches of bifurcation diagrams associated with the vertical paths (dashed lines) in the stability regions on the left. Solid lines correspond to stable states and dashed lines to unstable states.* T *stands for trivial state.* (a) $\delta = \delta' = 1$*,* $s = s' = 1$*,* $\rho = 0.4$*,* $\rho' = 0.67$*,* $\nu = 0.62$*, and* $\nu' = -1.02$*.* (b) $\delta = 5$*,* $\delta' = 0.5$*,* $s = s' = 1$*,* $\rho = 1.5$*,* $\rho' = 0.5$*, and* $\nu = -\nu' = 0.05$*. In* (a) *the traveling-wave branch that arises at* TW^+ *is not shown since we do not consider stability properties of traveling solutions.*

on the nonlinear coefficients, all four branches of pure modes and mixed modes (S_1, S_2, $S_\pm$) or just the pure modes (S_1,S_2) may arise at the intersection point of the neutral curves. Because we are considering $s = s' = 1$, S_1 and S_2 bifurcate supercritically. For $k = 0$ the eigenvalues $\lambda_\pm^{(1,2)}$ (cf. (3.6), (3.7)) determining the stability of S_1 and S_2 take the simpler form

$$(3.14)\qquad \begin{aligned} &\mathrm{L}_1 : \lambda_\pm^{(1)} = \mu(1-\rho') \pm \left\{\nu' \, |\mu|^{n/2}\right\}_{m=2} \qquad \text{for } S_1,\\ &\mathrm{L}_2 : \lambda_\pm^{(2)} = \mu(1-\rho) \qquad \text{for } S_2. \end{aligned}$$

Near onset (i.e., $0 < \mu \ll 1$) S_1 and S_2 are stable if $1 - \rho' < 0$ and $1 - \rho < 0$, respectively.

In the case $m = 2$, the stability of S_1 can be modified at larger values of μ by the resonant terms. Near onset the amplitudes of the mixed modes are approximated by

$$S_\pm: \quad R_1^2 = \frac{1}{1-\rho\rho'}\left(\mu(1-\rho) \mp (\nu - \nu'\rho)o(\mu^2)\right), \qquad (3.15)$$
$$R_2^2 = \frac{1}{1-\rho\rho'}\left(\mu(1-\rho') \mp (\nu - \nu'\rho)o(\mu^2)\right).$$

Note that small-amplitude mixed modes $S_\pm$ can therefore exist only if $(1-\rho)(1-\rho') > 0$. As discussed above, when $\rho\rho' < 1$, the stability of $S_\pm$ is controlled by the phase eigenvalue $e_\pm$ of (3.11). Upon substituting (3.15) into (3.11), we find that

$$\mathrm{sign}(e_\pm) = \mathrm{sign}\left\{\mp\frac{\mu}{1-\rho\rho'}\left(n\nu(1-\rho') + m\nu'(1-\rho)\right)\right\}. \qquad (3.16)$$

The results for $\Delta\mu = \gamma = 0$ and $s = s' = 1$ may now be summarized:

1. If $1-\rho > 0$ and $1-\rho' > 0$, both pure and mixed modes bifurcate from the trivial state as the forcing μ is increased. The two pure modes are both unstable, while one of the mixed modes, either S_+ or S_-, is stable. This case is illustrated in the diagram of Figure 3(a); in this example $\mathrm{sign}(e_\pm) = \mp 1$ (see (3.16)), implying that S_+ is stable and S_- is unstable at onset.
2. If $1-\rho < 0$ or $1-\rho' < 0$, only the two pure modes are present at onset. When $1-\rho\rho' > 0$, S_1 is stable if $1-\rho' < 0$ and S_2 is stable if $1-\rho < 0$. In particular, a bistable situation is permitted. In the example shown in Figure 3 only S_2 is stable because $1-\rho' > 0$.

Figure 4 presents two possible unfoldings of the degenerate diagrams shown in Figure 3. In both cases the curves $\alpha = 0$ and $\beta = 0$ intersect twice because $\chi > 0$. In Figure 4(a) the pure mode S_2 appears first ($\Lambda > 0$) and is stable. With increasing μ, however, perturbations in the direction of the other mode, A_1, become increasingly important, eventually destabilizing S_2 at L_2 in favor of the mixed modes $S_\pm$. Since $R_1 \ll R_2$ in the vicinity of L_2 and $\nu > 0$, S_+ is the stable mixed mode (cf. (3.11)). In Figure 4(b) we have $\Lambda < 0$, and the roles of S_1 and S_2 are switched. The mode S_1, which is now stable at onset, is destabilized by a perturbation in the direction of A_2 at L_1^-, generating the mixed state S_-. It undergoes a second bifurcation involving a perturbation in the direction of A_2 at L_1^+, leading to the mixed mode S_+. In this case S_- is stable, not S_+. In contrast with Figure 4(a), however, the pure mode S_2 is ultimately stabilized at large μ since $\rho > s' = 1$. Therefore the cross-interaction term $\rho|A_2|^2 \equiv \rho(\mu + \Delta\mu)/s'$ in (2.4), (2.5) dominates the linear growth rate μ of A_1 for large μ and suppresses the perturbations in the direction of A_1.

Figure 5 shows diagrams for the case where the curves $\alpha = 0$ and $\beta = 0$ do not cross ($\chi < 0$). In Figure 5(a) the pure mode S_2 bifurcates first, while S_1 does so in Figure 5(b). The subsequent bifurcations that these states undergo are basically of the same type as those in Figure 4. For example, in Figure 5(a) the mode S_2 becomes unstable to the mixed modes at L_2, while S_1 eventually gains stability at large μ because $\rho > s'$. A feature of Figure 5(b) that is not present in the previous diagrams is the saddle-node bifurcation on the S_+ branch. The appearance of this bifurcation is not specific to the case $\chi < 0$ but depends on the nonlinear coefficients, in particular on the resonance coefficients ν and ν' (see (3.5)).

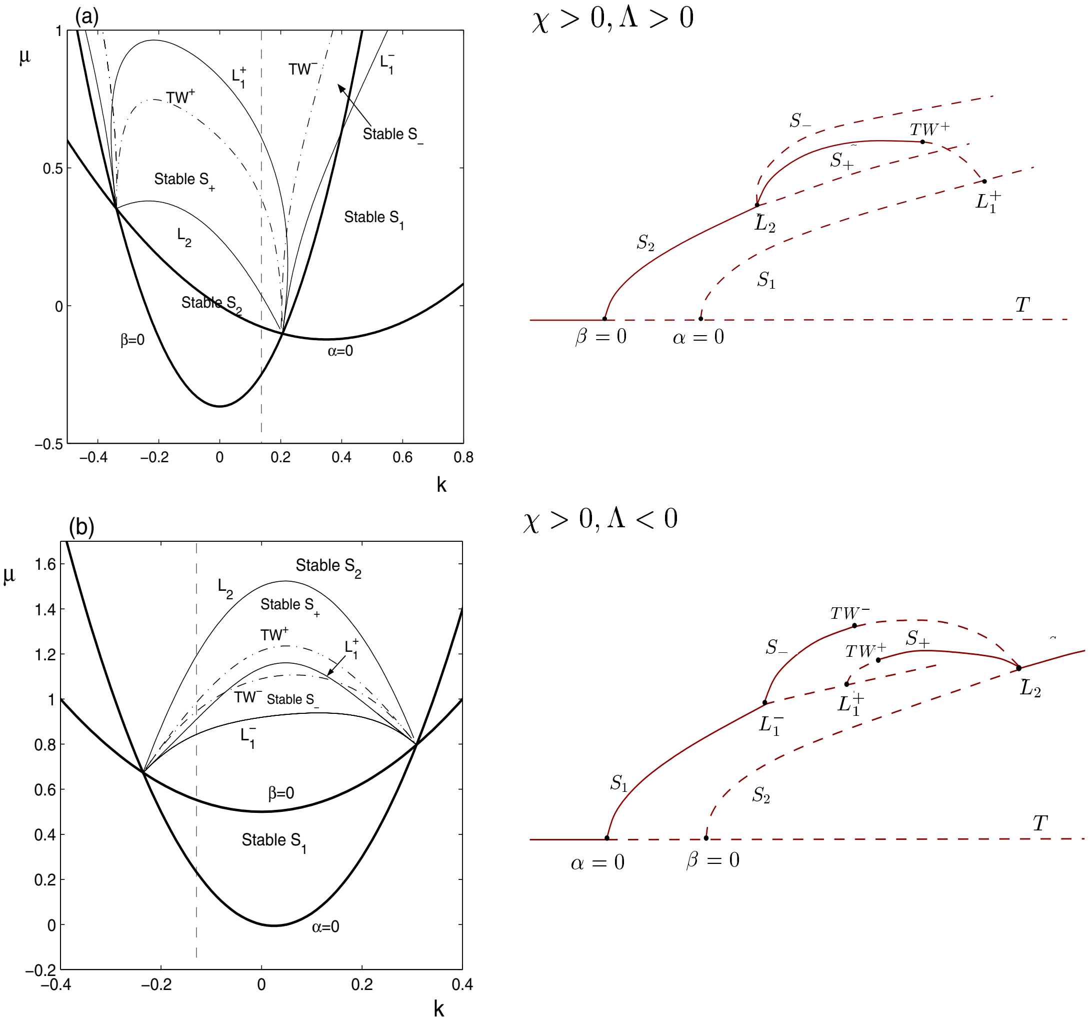


Figure 4. *Stability regions of the pure modes and mixed modes for the resonance* $2:5$. *Figures on the right are bifurcation diagrams associated with the vertical paths (dashed lines) in the stability diagrams on the left.* T *stands for trivial state* $A_1 = A_2 = 0$. (a) *Same parameters as in Figure* 3(a) *but* $\Delta\mu = 0.366$ *and* $\gamma = 0.7$. (b) *Same parameters as in Figure* 3(b) *but* $\Delta\mu = -0.5$ *and* $\gamma = 0.5$. *The branch of traveling waves that arises from* TW^- *in* (a)*, or that (possibly) connects the* S_+ *with the* S_- *solutions in* (b)*, is not shown.*

A comment regarding the validity of the phase diagrams is in order. For the weak resonances we are considering in this paper the resonant terms proportional to ν and ν' are of higher order in the amplitudes. Consequently, they have to be considered as perturbation terms. Since they are responsible for the splitting of the lines L_1 and TW into $L_1^{\pm}$ and $TW^{\pm}$, only those aspects of the results presented in Figures 3, 4, and 5 that persist as this splitting becomes small are expected to hold systematically. Thus, the transition to traveling waves at $k = 0$ in Figure 4(b) is robust, while that at $k = 0$ in Figure 4(a) is shifted to ever larger values of μ as the resonant terms become weaker. Similarly, the saddle-node bifurcation of S_+ in Figure 5(b) is not robust. It disappears through a sequence of bifurcations that involves a

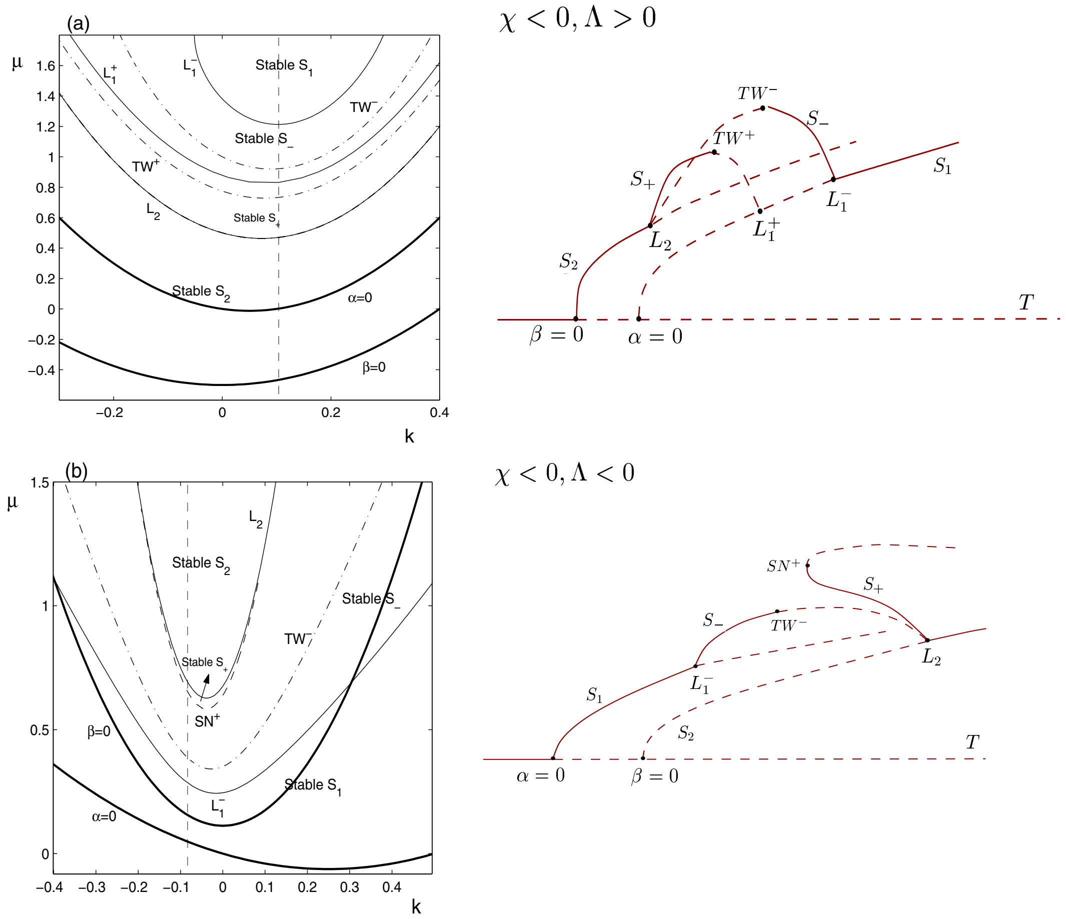


Figure 5. *Stability regions of the pure modes and mixed modes for the resonance* $2:5$. *Figures on the right are sketches of bifurcation diagrams associated with the vertical paths (dashed lines) in the stability diagram on the left.* T *stands for trivial state* $A_1 = A_2 = 0$. (a) $\Delta\mu = 0.5$, $\gamma = 0.5$, $\delta = 5$, $\delta' = 0.5$, $s = s' = 1$, $\rho = 0.5$, $\rho' = 1.5$, *and* $\nu = -\nu' = 0.085$. (b) $\Delta\mu = -0.1125$, $\gamma = 0.5$, $\delta = \delta' = 1$, $s = s' = 1$, $\rho = 1.5$, $\rho' = 0.5$, $\nu = 0.5$, *and* $\nu' = -1.01$. *The branch of traveling waves that (possibly) connects the* S_+ *with the* S_- *solutions in* (a), *or that arises from* TW^- *in* (b), *is not shown.*

merging of the branch S_+ with S_1. For very small $|\nu|$ and $|\nu'|$ the branch S_+ merges with S_1 close to L_1^-.

4. Side-band instabilities.

We now turn to the analytical core of this paper, the stability of the spatially periodic solutions with respect to side-band instabilities. Such instabilities are expected to destroy the periodicity of the solutions and provide a possible connection with quasi-periodic patterns. We therefore consider perturbations of the periodic solutions in the form

$$A_1 = (1 + a_1(X,T))R_1 e^{i(kX+\hat{\phi}_1)}, \quad A_2 = (1 + a_2(X,T))R_2 e^{i((nk/m)X+\hat{\phi}_2)}, \tag{4.1}$$

where $a_j(X,T) = (a_j^+(T)e^{iQX} + a_j^-(T)e^{-iQX})$ with $Q \neq 0$ and $j = 1,2$. Note that the perturbation wavenumber Q is measured relative to the deviation wavenumbers k and nk/m. The linear stability of the pure modes $S_{1,2}$ and the mixed modes $S_\pm$ is calculated by inserting the ansatz (4.1) into (2.4), (2.5) and linearizing in $a_j^\pm$. The details of these calculations are given in Appendix A.

4.1. Pure modes $S_{1,2}$. The linearized system for the perturbations associated with the pure modes $S_{1,2}$ separates into two uncoupled 2×2-blocks, which allows the stability of each pure mode to be calculated analytically. These two blocks can be associated with longwave and shortwave instabilities, respectively. The block corresponding to the longwave instability contains the eigenvalue related to spatial translations (i.e., phase modulations). The other (shortwave) block describes the evolution of amplitude perturbations in a direction transverse to the relevant pure-mode subspace. We find that in addition to the instabilities discussed in section 3 the destabilization of the pure modes can occur by longwave (Eckhaus) or shortwave instabilities.

In the case of S_1, a straightforward calculation yields

$$E_1 : \mu - \delta k^2 - \gamma k - (\gamma + 2k\delta)^2/2\delta = 0, \tag{4.2}$$

$$\Gamma_1 : (s\beta - \rho'\alpha)/s + \delta'(nk/m)^2 + \left\{\nu'^2|\alpha/s|^n/4\delta'(nk/m)^2\right\}_{m=2} = 0, \tag{4.3}$$

where E_1 is the Eckhaus curve and Γ_1 is the stability limit associated with shortwave perturbations. If the curve Γ_1 is crossed, S_1 undergoes a steady-state bifurcation with a perturbation wavenumber Q given by

$$Q^2 = \left(\frac{n}{m}\right)^2 k^2 - \left\{\nu'^2\left|\frac{\alpha}{s}\right|^n \left(\frac{m}{n}\frac{1}{2\delta' k}\right)^2\right\}_{m=2}. \tag{4.4}$$

When $m > 2$ the bracketed terms in (4.3) and (4.4) are absent. The destabilizing eigenfunction then takes the simple form $(a_1, a_2) \propto (0, e^{-ik\frac{n}{m}X})$ (see (4.1) and (A.4) of Appendix A) and allows a correspondingly simple physical interpretation: the wavenumber of the destabilizing mode is the wavenumber at the band center of the other mode S_2 (see (4.1)). This is no longer the case for $m = 2$ because the resonance terms (which are linear in A_2) affect the stability of S_1. One consequence of this is that the destabilizing mode is composed of two wavenumbers: $k\frac{n}{m} \pm Q$ with Q given by (4.4). Moreover, it is possible to have one or more nonzero wavenumbers k^*, say, for which the perturbation wavenumber Q vanishes. At these points $(k,\mu) = (k^*,\mu^*)$ the curve Γ_1 merges with the curves L_1^+ (or L_1^-) describing stability under homogeneous perturbations.

In the case of S_2 we find

$$E_2 : \mu + \Delta\mu - 3\delta'(nk/m)^2 = 0, \tag{4.5}$$

$$\Gamma_2 : (s'\alpha - \rho\beta)/s' + (\gamma + 2\delta k)^2/4\delta = 0, \tag{4.6}$$

where E_2 is the Eckhaus curve and Γ_2 is the stability limit for shortwave instabilities. As Γ_2 is crossed the pure mode S_2 undergoes a steady-state bifurcation with perturbation wavenumber

Q given by

$$Q = \pm|k + \gamma/2\delta|. \tag{4.7}$$

The associated eigenfunction is of the form $(a_1, a_2) \propto (e^{-i(k+\frac{\gamma}{2\delta})X}, 0)$ (see (4.1) and (A.9)) and, as in the case of S_1, points to a simple interpretation: the mode that destabilizes S_2 lies at the band center of S_1; i.e., its wavenumber is $\gamma/(2\delta) = \hat{\gamma}$ (cf. (4.1) and Appendix A; see also Figure 1). This result holds for all resonances except $m : n = 1 : 2$, in which case the linear stability of the pure mode S_2 is affected by the resonance terms (as S_1 was when $m = 2$).

Stability results for the pure modes $S_{1,2}$ are shown in Figures 6 and 7. In Figures 6(b),(d) and 7(b) the parameters are as in Figures 4(a),(b) and 5(a), respectively, while in the diagrams of Figures 6(a),(c) and 7(a) we depart from those parameters only in setting $\gamma = 0$. The inclusion of these last three cases allows us to gauge how much the side-band instabilities are affected by the detuning from perfect resonance.

Note that due to the large ratio δ/δ' the detuning of $\gamma = 0.5$ has only a small effect on the neutral curve of A_1 ($\alpha = 0$). Observe that in Figure 6(a),(b) the pure mode S_2 is stable within the region bounded by the curves E_2 and Γ_2, while S_1 is everywhere unstable due to shortwave instabilities. In contrast, Figure 6(c),(d) depicts a situation with stable regions for both pure modes. Note that the pure mode S_2 experiences only shortwave instabilities as the forcing is decreased while the pure mode S_1 can lose stability to either longwave or shortwave perturbations, provided the wavenumber k is not within the interval defined by the merging points of Γ_1 and L_1^- (see Figures 6(c),(d)); over that interval the instability suffered by S_1 is to perturbations preserving the periodicity of S_1 and gives rise to (steady) mixed modes. The parameters of Figure 7 do not allow for codimension-two points (intersection of the curves $\alpha = 0$ and $\beta = 0$). Despite this difference, the stability results for S_2 are qualitatively similar to those of Figure 6. The effect of γ on S_1, however, is more striking than in Figure 6(c),(d). In particular, S_1 can undergo a shortwave steady-state instability (along Γ_1) when $\gamma = 0.5$ but not when $\gamma = 0$.

4.2. Mixed modes $S_\pm$.

The stability of the mixed modes $S_\pm$ is governed by four eigenvalues which must be determined numerically. Two of these are associated with amplitude modes and are always real. The remaining two may be real or complex and are related, respectively, with the translation mode and with the relative phase between the two modes A_1 and A_2. Throughout this stability analysis we focus on the influence of the detuning parameter γ on the (side-band) instabilities of $S_\pm$. Furthermore, because of the invariance of (2.4), (2.5) under the transformation

$$\gamma \to -\gamma, \quad X \to -X,$$

we may take $\gamma \geq 0$ without any loss of generality.

The results of this stability analysis are shown in Figure 6, which depicts a situation with two codimension-two points, and Figure 7, which presents a case with no codimension-two points. Since this difference seems to have only a minor effect on the stability of $S_\pm$, we focus our discussion on the case of Figure 6 (i.e., two codimension-two points).

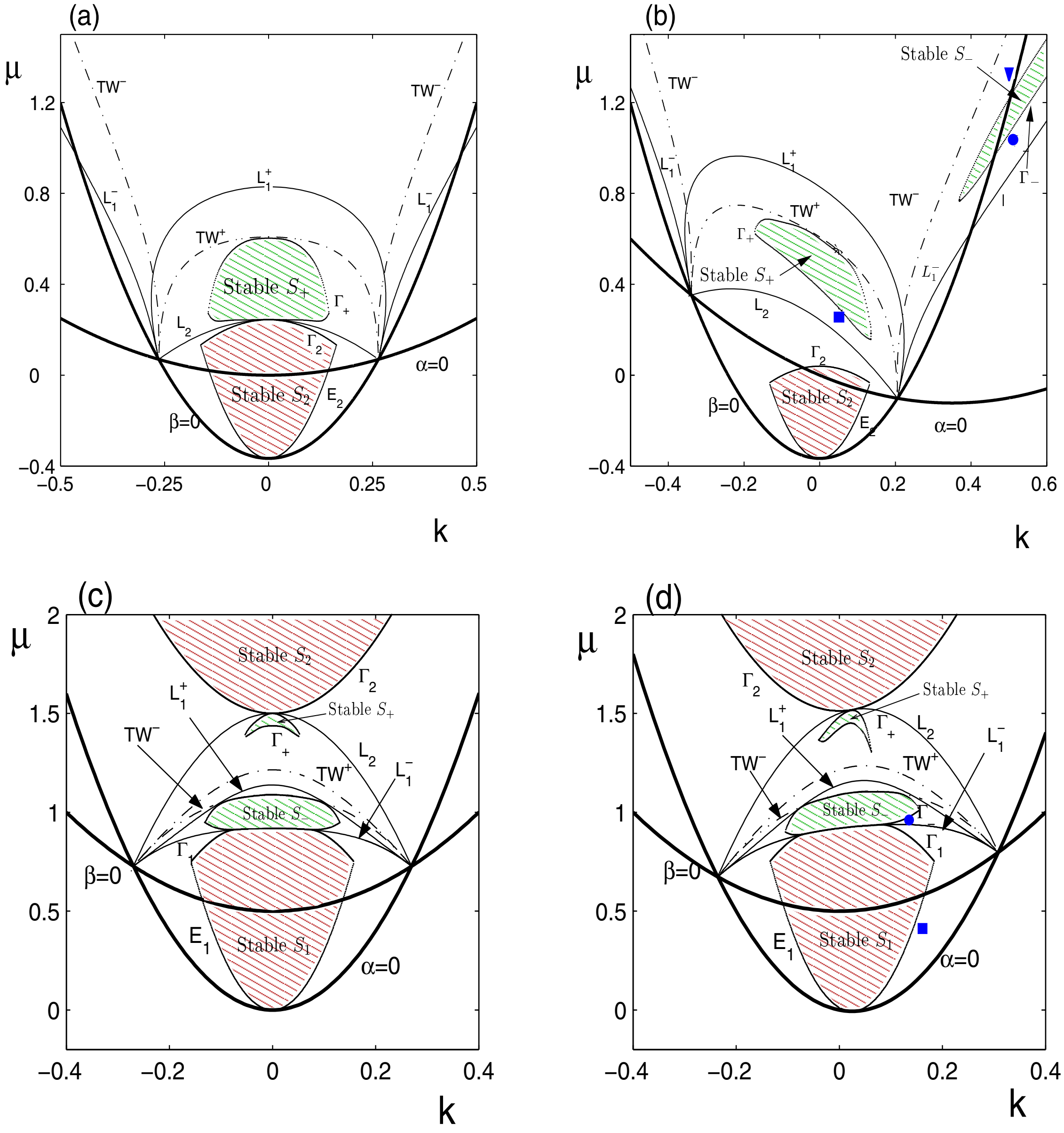


Figure 6. *Stability regions of the pure modes $S_{1,2}$ and the mixed modes $S_\pm$ for the resonance $2:5$ and the following parameter sets:* (a) $\Delta\mu = 0.366$, $\delta = \delta' = 1$, $s = s' = 1$, $\rho = 0.4$, $\rho' = 0.67$, $\nu = 0.62$, $\nu' = -1.02$, *and* $\gamma = 0$*;* (b) *as in* (a) *but* $\gamma = 0.7$ *(cf. Figure* 4(a)*);* (c) $\Delta\mu = -0.5$, $\delta = 10$, $\delta' = 0.5$, $s = s' = 1$, $\rho = 1.5$, $\rho' = 0.5$, $\nu = -\nu' = 0.05$, *and* $\gamma = 0$*;* (d) *as in* (c) *but* $\gamma = 0.5$ *(cf. Figure* 4(b)*). The symbols in* (b) *and* (d) *indicate the parameters used in the numerical simulations described in section* 5.

4.2.1. Mixed mode S_+.

When $\gamma = 0$ (Figure 6(a),(c)), the mixed mode S_+ becomes unstable to short-wavelength perturbations on the part of Γ_+ closer to L_2, while a longwave stability analysis captures the part of Γ_+ closer to TW^+ . Both instabilities are of steady-

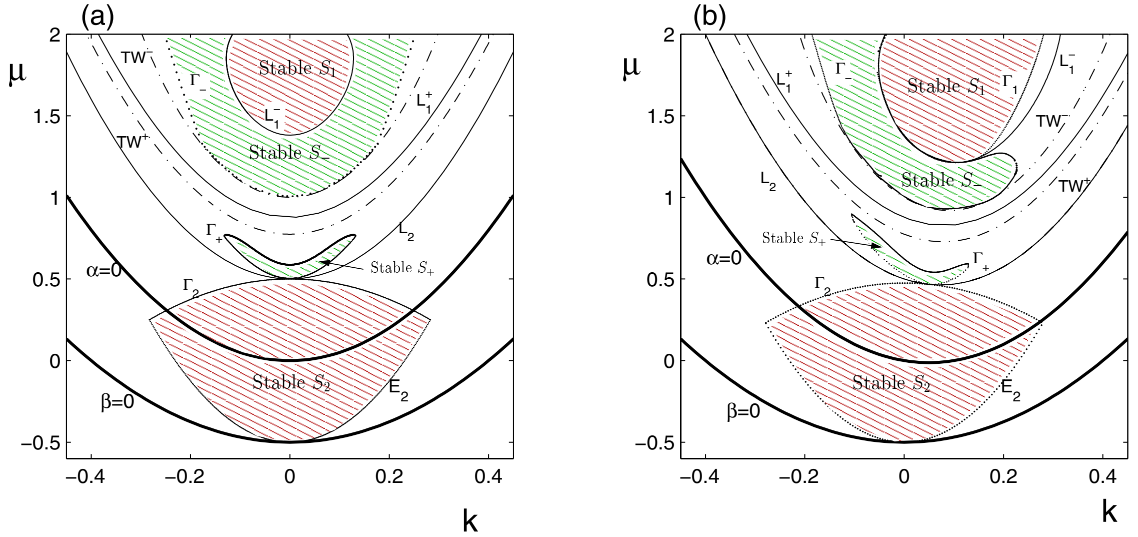


Figure 7. *Stability regions of the pure modes $S_{1,2}$ and the mixed modes $S_\pm$ for the resonance $2:5$ and the following parameter sets: (a) $\Delta\mu = 0.5$, $\gamma = 0$, $\delta = 5$, $\delta' = 0.5$, $s = s' = 1$, $\rho = 0.5$, $\rho' = 1.5$, $\nu = -\nu' = 0.085$.* (b) *as in*(a) *but $\gamma = 0.5$ (cf. Figure 5*(a)*).*

state type with the one close to TW^+ breaking the reflection symmetry of the pattern. This is expected to lead to a drift of the pattern; additional calculations for other parameter sets (not shown) suggest that this situation is characteristic of S_+ and $\gamma = 0$.

When $\gamma \neq 0$ the mixed mode S_+ displays both steady-state and Hopf bifurcations. The two relevant eigenvalues are associated with the translation mode and the symmetry-breaking mode, respectively, with the latter eigenvalue going to zero at TW^+. The change in character of the instability along Γ_+ is illustrated in Figure 8. Figures 8(b),(c) give the growth rates of the dominant modes as Γ^+ is crossed at specific points marked in Figure 8(a). At points 1–5 the bifurcation is steady, but as one continues in a counterclockwise direction (points 4–6) the eigenvalues merge and become complex over an interval of Q-values. This interval containing complex conjugate eigenvalues continues to grow, leading eventually to an oscillatory instability superseding the steady one. A codimension-two point therefore exists near $(k, \mu) = (-0.155, 0.581)$ at the transition from steady-state to Hopf bifurcation; at this point there are two unstable modes, at different Q, one steady and one oscillatory. As one goes still further in the counterclockwise direction, the Q-band over which the eigenvalues are complex reaches $Q = 0$ and the oscillatory instability becomes a long-wave instability. This interaction was studied in the context of Taylor vortex flow [43] and has been found in many others problems [41, 21].

Similar transitions occur if μ and γ are varied (rather than μ and k). We illustrate this in Figure 9 by plotting the growth rate of the most unstable perturbations after crossing, at fixed $k = -0.1$, the upper part of Γ_+ which is close to TW^+ (Figure 9(a)), and the lower part of Γ_+ which is close to L_2 (Figure 9(b)). Along the upper part of Γ_+ the mixed mode S_+ typically sees a shortwave oscillatory instability; for small γ, S_+ can also lose stability

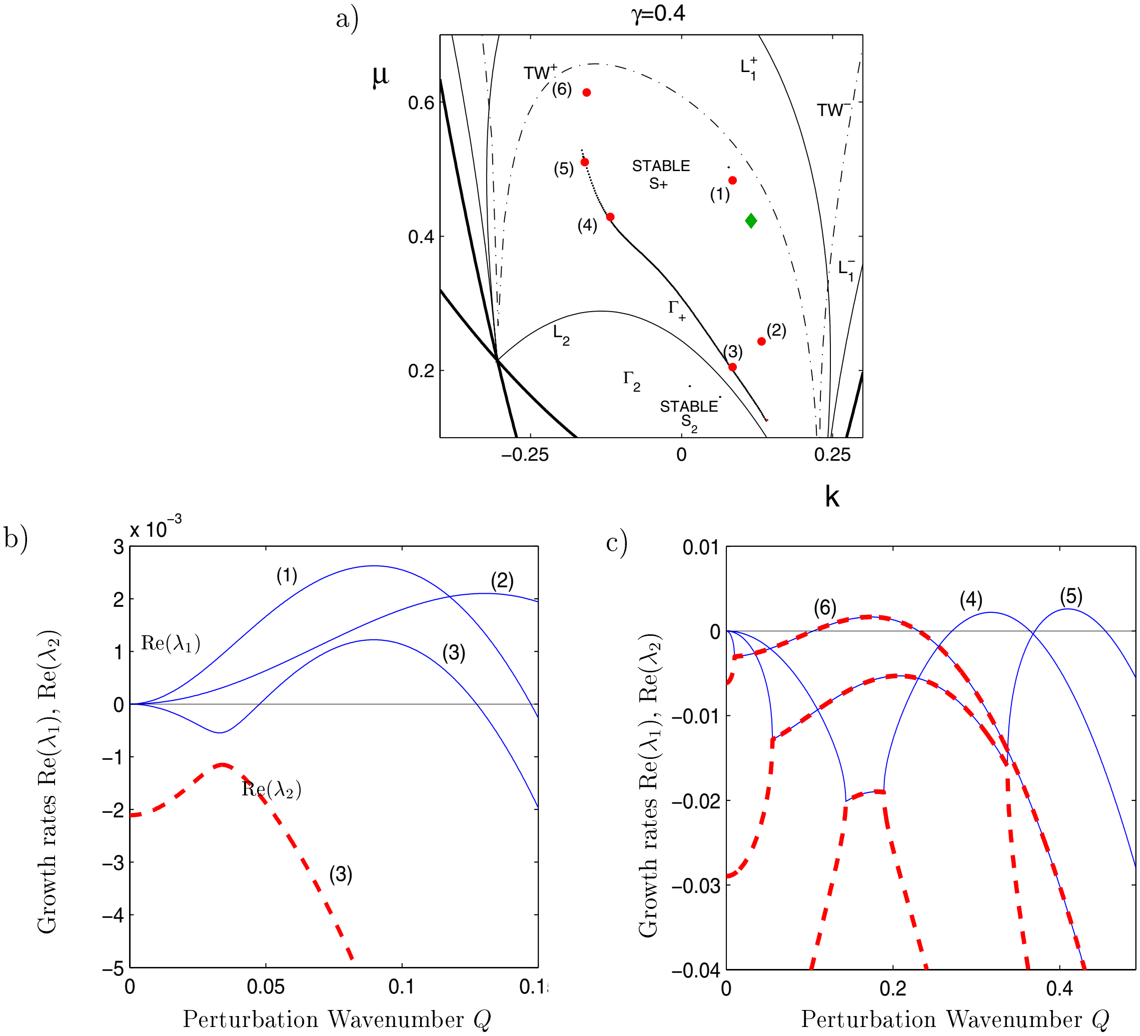


Figure 8. *Growth rates of the two most unstable perturbations of S_+ when Γ_+ is crossed at the points indicated in the upper figure:* $\mathrm{Re}(\lambda_1)$ *(solid-thin line) and* $\mathrm{Re}(\lambda_2)$ *(dashed-thick);* λ_1 *and* λ_2 *are complex where there is a single curve, i.e., where* $\mathrm{Re}(\lambda_1) = \mathrm{Re}(\lambda_2)$*. We use* $\gamma = 0.4$ *with the remaining parameters as in Figure* 6(b).

to longwave perturbations. The lower part of Γ_+ is characterized by either steady-state or oscillatory instability of shortwave type (see Figure 9(b)). There is again a codimension-two point, near $(\gamma, \mu) \simeq (1.035, 0.47)$, where Hopf and steady-state bifurcations coalesce. For γ less (greater) than this critical value the bifurcation is steady (oscillatory). The route to the oscillatory instability is as described above: with increasing γ the two modes associated with translation and reflection symmetry interact more and more, ultimately leading to a Hopf bifurcation.

The steady-state shortwave instability of S_+ (see Figure 9(b)) always appears in the vicinity of L_2. This proximity implies that the amplitude A_1 of the mixed mode is very small.

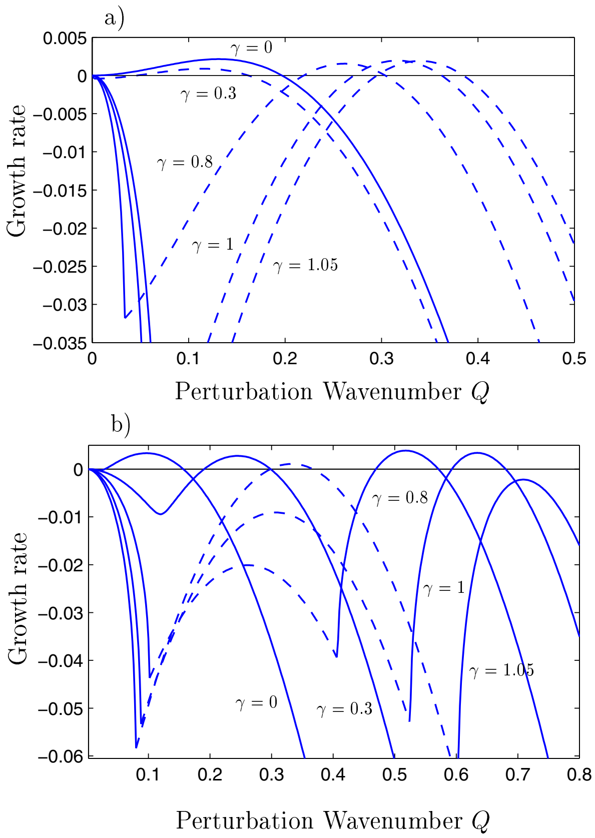


Figure 9. *Growth rate of the most unstable perturbation of S_+ when Γ_+ is crossed vertically at $k = -0.1$ for the indicated values of γ,* (a) *on the upper part of Γ_+ (close to TW^+) and* (b) *on the lower part (close to L_2). The remaining parameters are as in Figure* 6(a)–(b)*. Solid (dashed) lines correspond to real (complex) eigenvalues.*

Consequently, the behavior of S_+ in this region is similar to that of the pure mode S_2. The growing perturbations are predominantly in the direction of A_1, and the associated wavenumber Q is given in a first approximation by (4.7), i.e.,

$$Q \simeq \pm|k + \gamma/2\delta|. \tag{4.8}$$

As demonstrated in Figure 10, (4.8) provides a good approximation over the interval $0.1 < \gamma < 0.7$ when $|k| \leq 0.1$. It is interesting to note that, when γ is not too small, (4.8) applies independently of the resonance $m : n$. The robustness of this result is not unexpected since the resonance terms do not influence at first order the stability properties of the mixed mode S_+ (because $|A_1|$ is small). However, (4.8) ceases to apply when γ becomes very small because resonance effects start to play a significant role in the wavenumber selection of the linear response when crossing the lower part of Γ_+. In particular, for $k = 0$ and $\gamma \ll 1$ one can show

$$Q \simeq (\gamma^n \nu)^{1/4} \left[r_{20}(2s'/(ss' - \rho\rho'))^{3/4} \right]^{1/2} /2\delta + O(\gamma^{n+1}), \tag{4.9}$$

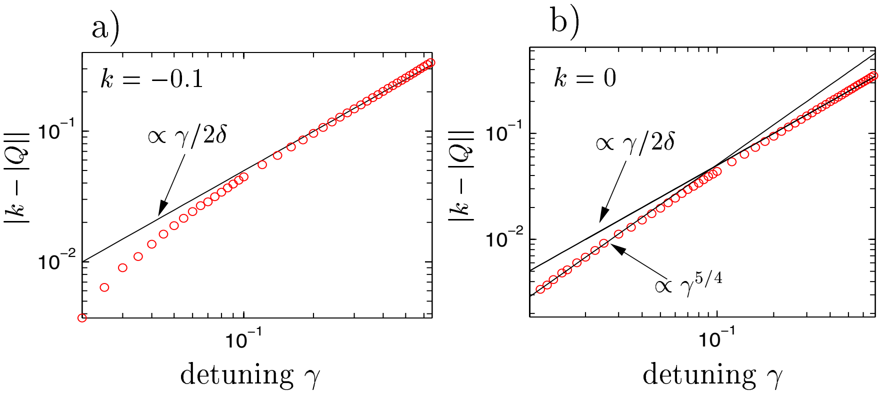


Figure 10. *Dependence of the perturbation wavenumber Q of the steady mode destabilizing S_+ on the detuning parameter γ for the resonance case $m : n = 2 : 5$. Parameters as in Figure* 6(a),(b) *with* (a) $k = -0.1$ *and* (b) $k = 0$. *Open circles correspond to the perturbation wavenumber Q calculated numerically from* (B.2), (B.3). *Solid lines correspond to the indicated approximations.*

where r_{20} denotes the amplitude of S_2 (with $k = 0$) on the curve L_2. The derivation of (4.9), given in Appendix B, relies on the fact that for $\gamma = 0$ the curve Γ_+ becomes tangent to the curve L_2 at $k = 0$. Figure 11 shows that (4.9) is in excellent agreement with the numerical results obtained directly from the characteristic equation (B.2) and applies to other weak resonances with $m : n \neq 2 : 5$. Note that for larger values of n (4.9) is valid over a wider range of γ values (since the error is $O(\gamma^{n+1})$) and the cross-over to the behavior (4.8) is shifted to larger values of γ.

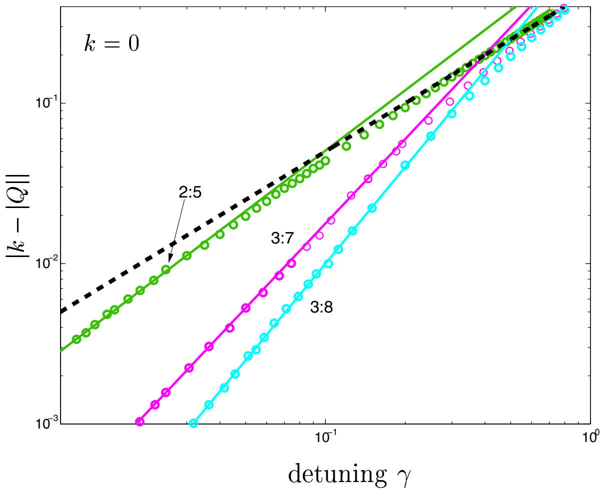


Figure 11. *Dependence of Q on the detuning parameter γ for the indicated resonances $m : n$ and $k = 0$. Remaining parameters are $\Delta\mu = 0.366$, $\delta = \delta' = 1$, $s = s' = 1$, $\rho = 0.4$, $\rho' = 0.67$, $\nu = 0.62$, and $\nu' = -1.02$. Open circles correspond to numerical results of* (B.2), (B.3) *and lines to the approximations* (4.8) *and* (4.9), *$Q \propto \gamma/2\delta$ (dashed) and $Q \propto \gamma^{n/4}$ (solid), respectively.*

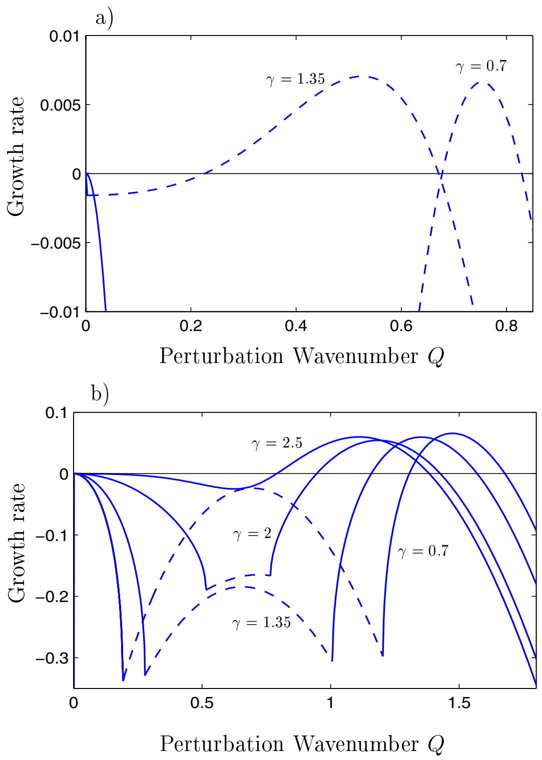


Figure 12. *Growth rate of the most unstable perturbation when Γ_- is crossed vertically at $k = 0.5$ for the indicated values of γ and the remaining parameters as in Figure* 6(a)–(b). (a) *On the upper part of Γ_- (close to TW^-) and* (b) *on the lower part (close to L_1^-). A solid line is used when the associated eigenvalue is real and a dashed line is used when it is complex.*

4.2.2. Mixed mode S_-.

As illustrated in Figure 6(a), the mixed mode S_- can be everywhere unstable when $\gamma = 0$. For the parameters of Figure 6(a), such a situation persists until $\gamma \approx 0.6$, when a stability region for S_- emerges. This stability region widens as γ is increased (see Figure 6(b)).

As in the case of S_+, the two most important eigenvalues for the stability of S_- are those associated with the translation and parity-breaking modes. The oscillatory instability results from an interaction between these two eigenvalues and, consequently, is found in the vicinity of the bifurcation set TW^-; the steady (shortwave) instability occurs near the initial bifurcation producing the mixed mode S_- from the pure mode (i.e., near L_1^-). Thus, the mixed mode S_- undergoes the same kinds of transitions described above for S_+. Depending on where Γ_- is crossed in the (k, μ)- or the (γ, μ)-plane, the instability may be either oscillatory or steady and either of long-wave or short-wave type.

Note that on the upper part of Γ_- in Figure 12(a), as γ is increased the two eigenvalues merge and become complex at ever smaller values of Q. In contrast, on the lower part of Γ_- raising γ increases the damping of the parity-breaking mode, thus discouraging its interaction with the translation mode; this behavior is opposite to what occurs with S_+. It is also interesting to note that variations in the detuning parameter γ do not induce a significant change in the perturbation wavenumber Q associated with the bifurcation. This is because the instability of S_- takes place close to L_1^-, where the destabilizing perturbations are (preferentially) in the direction of A_2. The perturbation wavenumber Q associated with the instability is given, to first approximation, by (4.4). Thus Q depends mainly on the resonance $m : n$ and only implicitly on the parameter γ (through the function α defined in (3.3)).

Concerns similar to those discussed in section 3 about the robustness of the stability diagrams when $\nu,\ \nu' \to 0$ apply here. Because the stability properties of the mixed modes with respect to side-band perturbations are essentially governed by the proximity to the bifurcations L_2, $L_1^\pm$, and $TW^\pm$, the stability diagrams in Figures 6(c),(d) and 7 are robust, as are the steady-state instabilities of S_+ that take place close to L_2 in Figures 6(a),(b). In the latter two cases, however, the oscillatory instabilities that $S_\pm$ undergo close to $TW^\pm$ will be shifted toward larger μ-values as ν and ν' get smaller.

5. Numerical simulations.

The main objective in this section is to investigate the nonlinear evolution of the side-band instabilities of the pure and mixed modes discussed in section 4. The results are obtained for a periodic domain of length L using a finite-difference code with Crank–Nicholson time-stepping. Each simulation of (2.4), (2.5) uses as an initial condition an unstable periodic steady state (i.e., a pure or mixed mode of the type discussed in section 3) with a small random perturbation. The final solutions obtained in this way are typically quasi-periodic although periodic solutions are also found. Here we use the term quasi-periodic to refer to any solution for which the wavenumbers of A_1 and A_2 are *not* in the ratio m/n. Therefore the full reconstructed field (2.2) will not be composed of wavenumbers in the ratio m/n; in general it will be quasi-periodic.

First we describe the evolution of system (2.4), (2.5) when the pure modes become unstable. If the instability is longwave, it affects only the excited amplitude and the solution remains in the pure-mode subspace. For example, in the case of S_1 (see Figure 13) the longwave instability involves A_1 but not A_2, which remains zero. As expected, the changes affect predominantly the phase of A_1 and evolve into a series of phase-slips. The final state, which is time-independent, is shown on the right part of Figure 13. In the case of S_2 (not shown) the longwave instability behaves analogously; it is now A_2 that changes, while A_1 remains zero.

Depending on the detuning parameter γ the transition from the pure modes to the mixed modes can also involve side-band instabilities. For the parameters corresponding to the stability diagram shown in Figure 7 the pure mode S_1 becomes unstable to the mixed mode S_- via a side-band instability for $\gamma \neq 0$ and $k > 0.15$ (cf. Figure 7(b)), while for $\gamma = 0$, i.e., when the two critical wavenumbers are resonant, the instability preserves the periodicity of the solution for all k (cf. Figure 7(a)). The final states that are reached after a small random perturbation has been applied to a pure-mode solution S_1 are depicted in Figure 14. The movies 60055_01.avi and 60055_02.avi show the corresponding temporal evolution of A_2. The sequence of phase-slips that is evidenced in the sharp peaks of the local wavenumber during the early stages of the evolution is a consequence of the random initial perturbation. After this transient the relevant perturbation mode starts to dominate. In the case $\gamma = 0.5$ that mode has a strongly modulated wavenumber reflecting the fact that the wavenumbers of A_1 and A_2 do not satisfy the resonance condition $m : n = 2 : 5$, resulting in a quasi-periodic reconstructed solution $u(x)$ (see also discussion leading to (5.1) below).

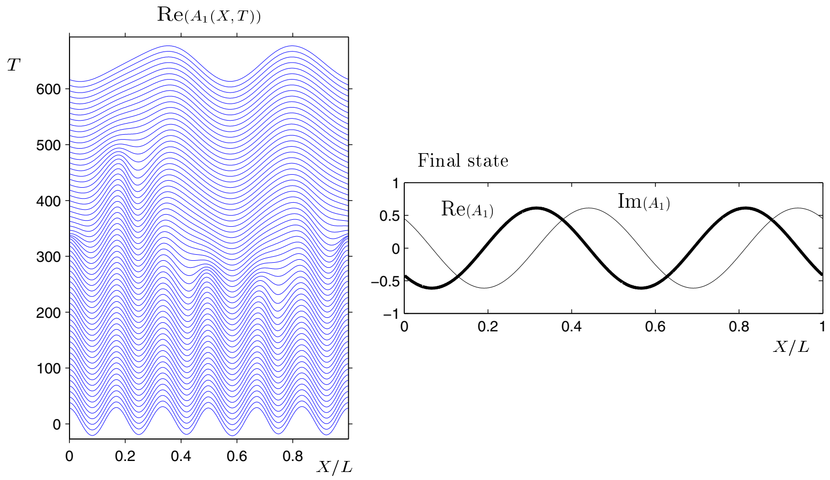


Figure 13. *Evolution of the longwave instability of the pure mode S_1. Space-time diagram of the real part of A_1 (left) and final state (right) with thick (thin) lines denoting the real (imaginary) part of A_1. Parameters as in Figure* 6(d) *with $k_1 = k = 0.15$, $\mu = 0.37$ (indicated by a square in Figure* 6(d) *and $L = 250$.*

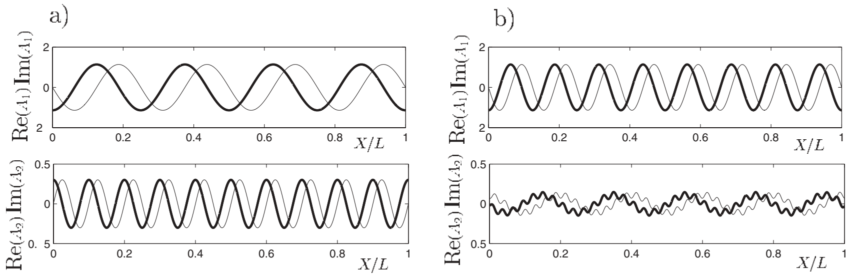


Figure 14. *Final state after a small random perturbation of the pure mode S_1. Parameters in* (a) *and* (b) *as in Figure* 7(a),(b)*, respectively, with $L = 250$ and $\gamma = 0$, $k = 0.1$, $\mu = 1.4$ in* (a) *and $\gamma = 0.5$, $k = 0.2$, $\mu = 1.4$ in* (b)*. The corresponding movies* (60055_01.avi *and* 60055_02.avi) *show the temporal evolution of A_2. Red and yellow lines show the real and imaginary part of A_2, respectively, white its magnitude $|A|$, and green the local wavenumber. Clicking on the above images displays the associated movies.*

The longwave instability of the mixed modes affects A_1 as well as A_2 (see Figure 15). As with the pure modes, the phase perturbations evolve into phase-slips, which change the wavenumber and lead eventually to a stationary state. In the case shown in Figure 15 phase-slips occur only in mode A_2. As in the mixed mode shown in Figure 14(b), the strong deviations of $A_{1,2}(x)$ from purely sinusoidal behavior, which are due to the resonant terms proportional to ν and ν', indicate that the reconstructed solution $u(x)$ is not periodic.

The mixed modes can undergo a shortwave steady instability as well. As discussed in section 4 it occurs when the stability limit Γ_+ (Γ_-) of S_+ (S_-) is close to the transition line L_2 (L_1^-). The instability is primarily in the direction of the mode with smaller amplitude: A_1 for Γ_+ near L_2 and A_2 for Γ_- near L_1^-. The resulting evolution of system (2.4), (2.5) is shown in Figure 16 for parameters near Γ_+. It confirms the expectation that at least initially only A_1 changes, but not A_2. The space-time diagram for $\mathrm{Re}(A_1(X,T))$ in Figure 16 shows how the growing perturbation determines the wavenumber of the final state; this change from the

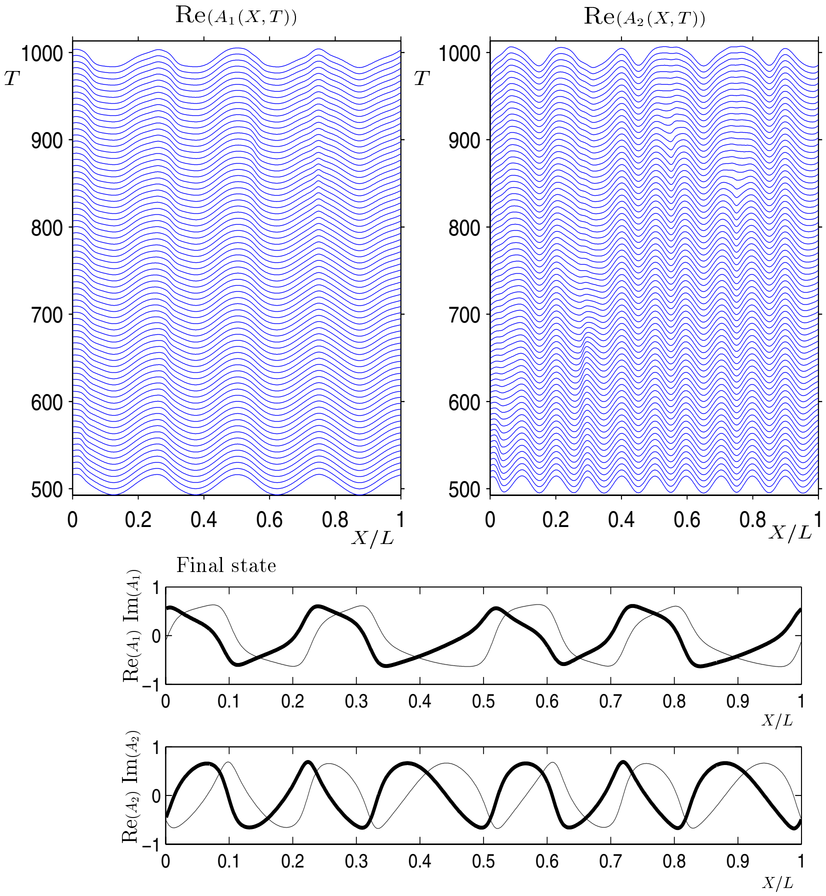


Figure 15. *Evolution of longwave instability of the mixed mode S_+. Shown are space-time diagrams of the real part of the amplitudes and the final state with thick (thin) lines denoting the real (imaginary) parts of A_1 and A_2, respectively. Parameters as in Figure 8*(a) *with $k = 0.1$, $\mu = 0.423$ (indicated by rhomb) and $L = 300$.*

initial wavenumber takes place via phase-slips. The wavenumber k_1 of A_1 in the final state is controlled by the detuning parameter γ in the manner predicted by the linear analysis: $k_1 \simeq \gamma/(2\delta)$. In other words, the number of maxima observed in A_1 is $N_1 \simeq L\gamma/(4\pi\delta)$ (see lower figures of Figure 16).

The evolution of A_1 and A_2 after a shortwave steady instability on Γ_- near L_1^- is shown in Figure 17. This time the amplitude A_1 remains nearly constant while A_2 changes. The destabilizing mode was shown in section 4 to be of the form $a_2^+(T)e^{i(k\frac{n}{m}+Q)X}+a_2^-(T)e^{i(k\frac{n}{m}-Q)X}$ with Q defined by (4.4). For the cases illustrated in Figure 17 we have $|a_2^+(T)| \ll |a_2^-(T)|$ and the wavenumber of the linearly unstable mode is approximately given by $k_2 \approx k\frac{n}{m} - |Q|$. Note that this wavenumber is essentially determined by the resonance terms and would be 0 without them (see (4.4)). In particular, for the example considered in Figure 17(a), where $\nu = 0.62$ and $\nu' = -1.02$, we have $|k_2| \simeq 0.144$, while in Figure 17(b) $\nu = -\nu' = 0.05$ and $|k_2| \simeq 0.018$; note that with regard to this effect both cases are analogous, except for the magnitude of the resonance coefficients.

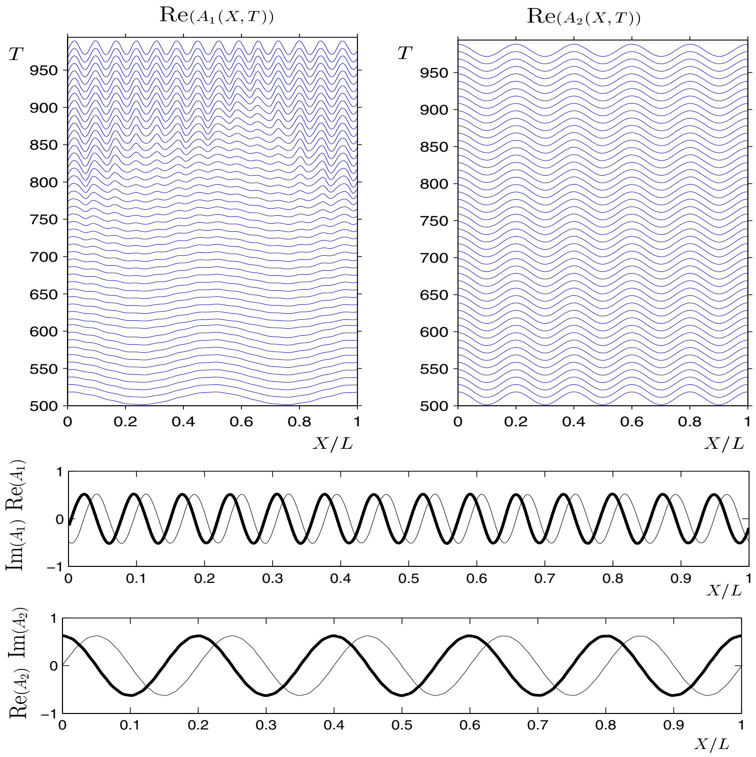


Figure 16. *Space-time diagram showing the real part of the amplitudes A_1 and A_2. The lower figures show the real (thick line) and imaginary (thin line) parts of A_1 and A_2 corresponding to the final stable stationary state. The parameters are as in Figure* 6(b) *(square) with $k = -0.05$, $\mu = 0.27$, and $L = 250$.*

In addition to the wavenumber k_2 determined by the linear stability analysis, Figure 17 reveals a second prominent wavenumber k_2', which is the result of the resonance term $\bar{A}_2^{m-1}A_1^n$. It acts as a driving term for A_2 and generates a mode with wavenumber k_2' that is determined by

$$nk_1 \pm (m-1)|k_2| = \pm|k_2'|, \tag{5.1}$$

where k_1 is the wavenumber associated with amplitude A_1. In the evolution shown in Figure 17 k_1 remains unchanged. The same mechanism is at work for the short-wave instability of S_+ (cf. Figure 16), which occurs near L_2. There the wavenumber modulation of A_1 is, however, negligible. It is driven by the resonance term $\bar{A}_1^{n-1}A_2^m$, which is proportional to the fourth power of the amplitude A_1 ($n = 5$), which is very small where S_+ branches off the pure mode S_2. This is not the case for the shortwave instability of S_- (Figure 17), which occurs near L_1^-. There it is A_2 that is small; but it enters the resonance term $\bar{A}_2^{m-1}A_1^n$ linearly ($m = 2$) and therefore the resonance term provides a relatively strong modulation of A_2. Substituting the final values of k_1 (same as the initial value) and k_2 (calculated above) into (5.1), one obtains $|k_2'| = 2.644$ in the case of Figure 17(a) and $|k_2'| = 0.63$ for Figure 17(b). For the numbers (N_1,N_2,N_2') of wavelengths associated with k_1, k_2, and k_2', respectively, this implies $(N_1, N_2, N_2') = (20, 6, 106) \simeq L/(2\pi)(0.5, 0.144, 2.644)$ for Figure 17(a) (where $L = 250$) and

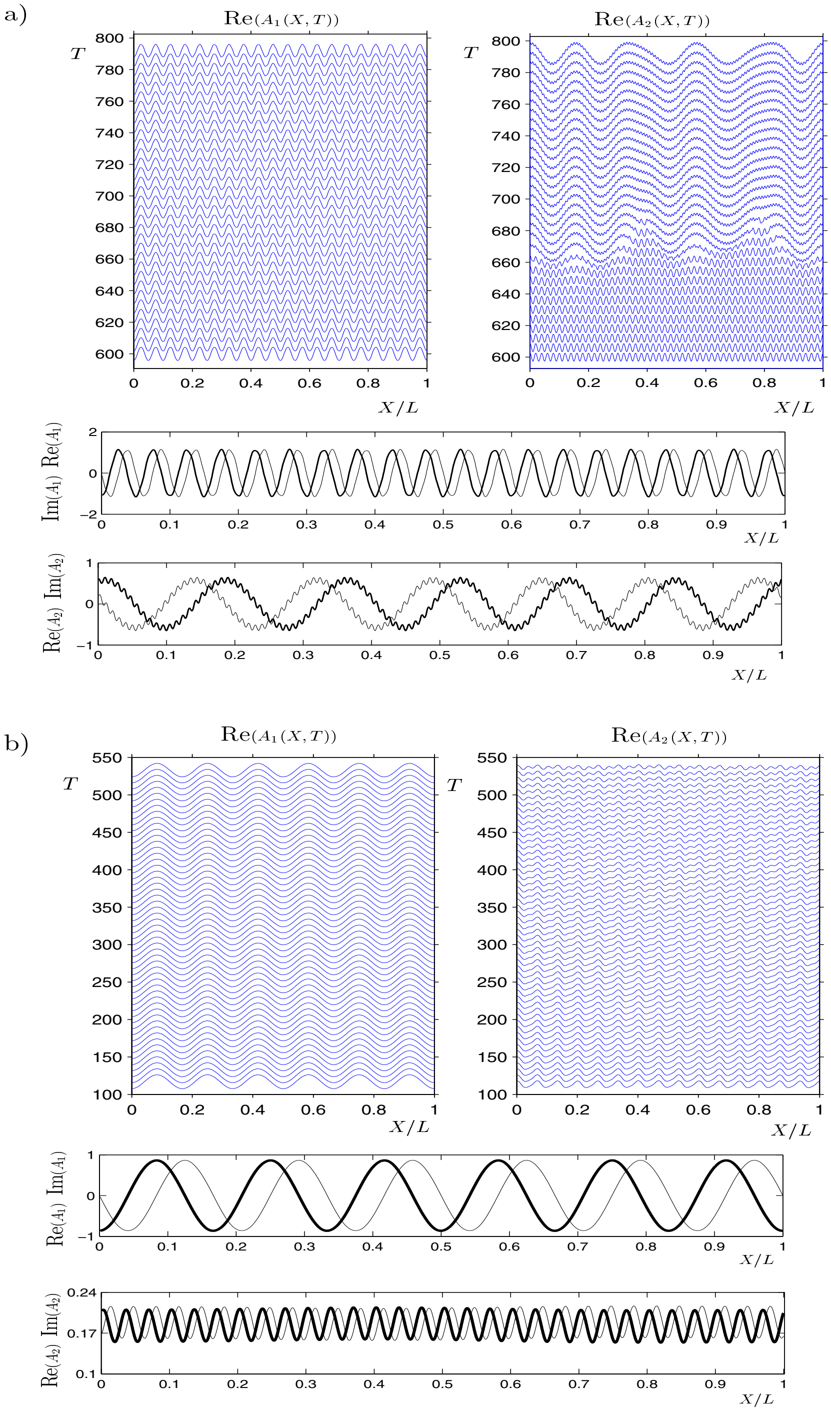


Figure 17. *Space-time diagram showing the real parts of* A_1 *and* A_2 *when mixed mode* S_- *undergoes a shortwave steady instability. The parameters in* (a) *are those of Figure* 6(b) *(circle) with* $k = 0.5$, $\mu = 1.05$, *and* $L = 250$. *Parameters in* (b) *are as in Figure* 6(d) *(circle) with* $k = 0.125$, $\mu = 0.95$, *and* $L = 305$.

$(N_1, N_2, N_2') = (6, 1, 31) \simeq L/(2\pi)(0.124, 0.02, 0.63)$ for Figure 17(b) (where $L = 305$). These results compare quite well with the results in Figure 17, where $(N_1, N_2, N_2') = (20, 6, 106)$ in Figure 17(a) and $(N_1, N_2, N_2') = (6, 0, 30)$ in Figure 17(b).

Qualitatively different is the evolution of the oscillatory instability of the mixed mode, as illustrated in Figures 18, 19, 20, and 21. Small perturbations initially evolve into standing-wave oscillations. They do not last for long, however, and decompose into left- and right-traveling disturbances, which then form localized low-wavenumber domains of waves drifting to the right and to the left, respectively. Such localized, propagating regions of "drift waves" are typical for parity-breaking instabilities because the extended drift waves are generically unstable at onset [43, 6, 19] and may become stable only for larger amplitudes [4]. The localized waves can be described using equations for the amplitude of the parity-breaking mode and the phase of the underlying pattern and can arise stably when the parity-breaking bifurcation is subcritical [9, 22] as well as when it is supercritical [43, 6]. They are related to the solitary modes observed in experiments in directional solidification [46, 18] as well as in Taylor vortex flow [49] and in simulations of premixed flames [4]. In these systems the parity-breaking bifurcation arises from a $1 : 2$ mode interaction [28, 43] rather than the $2 : 5$-resonance considered here.

The localized drift waves do not always persist. In the simulation shown in Figure 18 a sequence of phase-slips occurs in mode A_2 at the trailing end of the localized drift wave. The phase-slips substantially reduce the wavenumber in the growing domain between the location where the localized drift wave first is created (at $x/L \approx 0.6$) and its trailing edge. Since no phase-slips occur in mode A_1, the wavenumbers of A_1 and A_2 are not resonant anymore in this growing domain, which results in a strong modulation of the wavenumber. Due to the periodic boundary conditions, the growing localized drift wave collides eventually with this domain and is absorbed by it. After that the pattern becomes stationary. The stationary domains with strongly modulated wavenumber do not always absorb the localized drift waves. Figure 19 and the movie 60055_03.avi show a case in which the localized drift waves at times are also reflected by the stationary domains or pass through them. We have not investigated which factors determine the outcome of the collisions. In the simulation shown in Figure 19 the system eventually evolves into a stationary state. It consists of relatively large domains in which the wavenumber ratio is very close to $2 : 5$ and small domains in which the wavenumbers are strongly modulated and not in this rational ratio. One may have expected that the localized stationary domains attract each other and eventually merge; but this was not observed. Presumably, the strong wavenumber oscillations, which are particularly visible in A_2, lock the domains in place. Thus, in the final, stationary state shown in Figure 19(b) one has a periodic pattern that is interrupted by small domains in which the pattern is not periodic. Stationary localized structures have been investigated in the case of two nonresonant Ginzburg–Landau equations (i.e., with no phase coupling between the two unstable modes) in terms of homoclinic and heteroclinic solutions (in space) [16]. The extent to which these patterns are related to our case is not clear, since the resonant terms may play a crucial role in fixing the ratio of the wavenumbers of the background state (Figure 19).

The transients and the resulting final state depend very sensitively on the amplitude of the initial perturbations of the periodic state. The relatively large perturbations used in Figure 19 lead to a large number of localized drift waves, which upon their collision generate localized

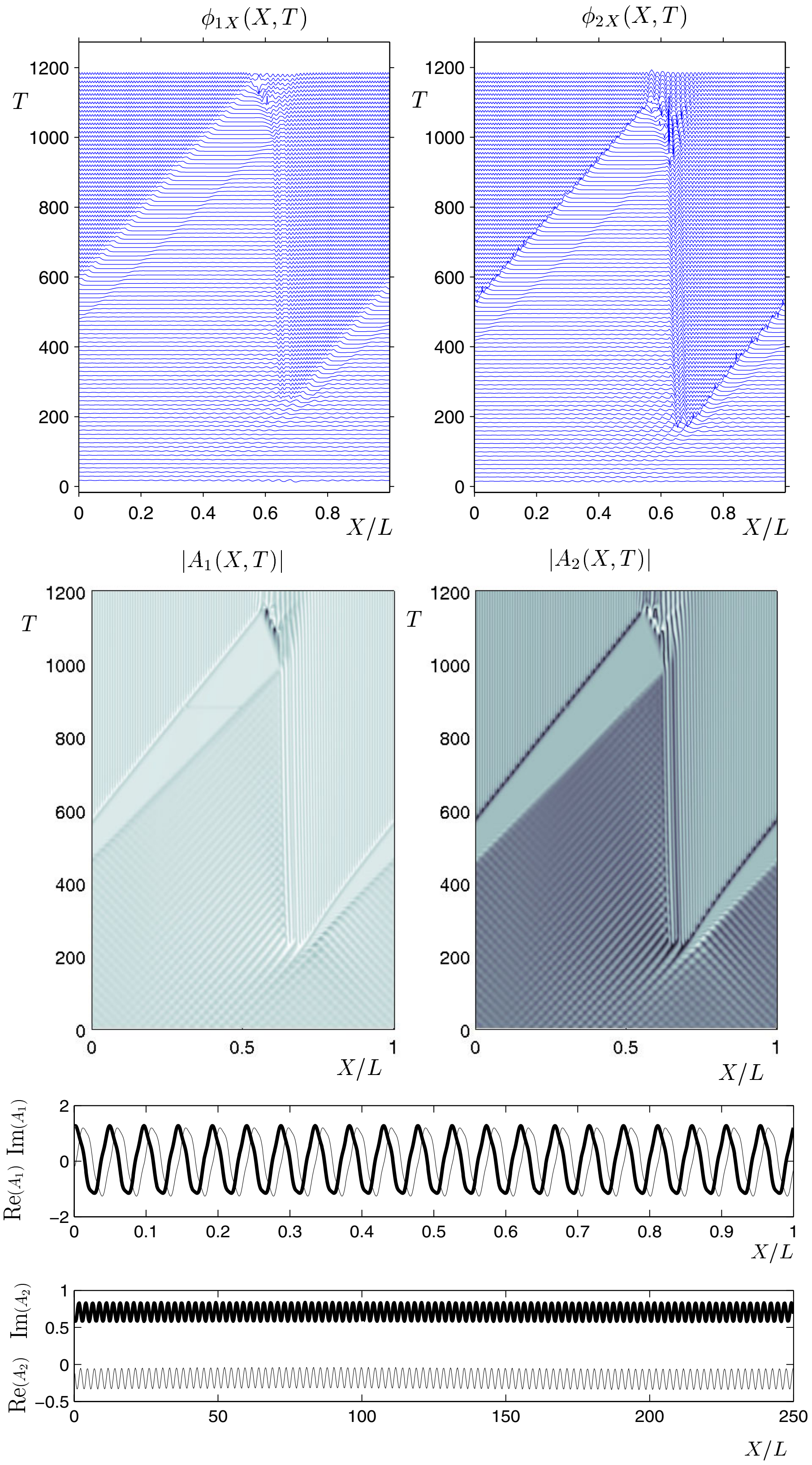


Figure 18. *Space-time diagram of the phase-gradient and the magnitude of the amplitudes A_1 and A_2. Lower figures: Real (thick) and imaginary (thin) part of A_1 and A_2 corresponding to the final stable state. Parameter as in Figure* 6(b) *(triangle) with $k = 0.5$, $\mu = 1.308$, and $L = 250$. In the gray-scale plots dark (light) stands for low (high) amplitudes.*

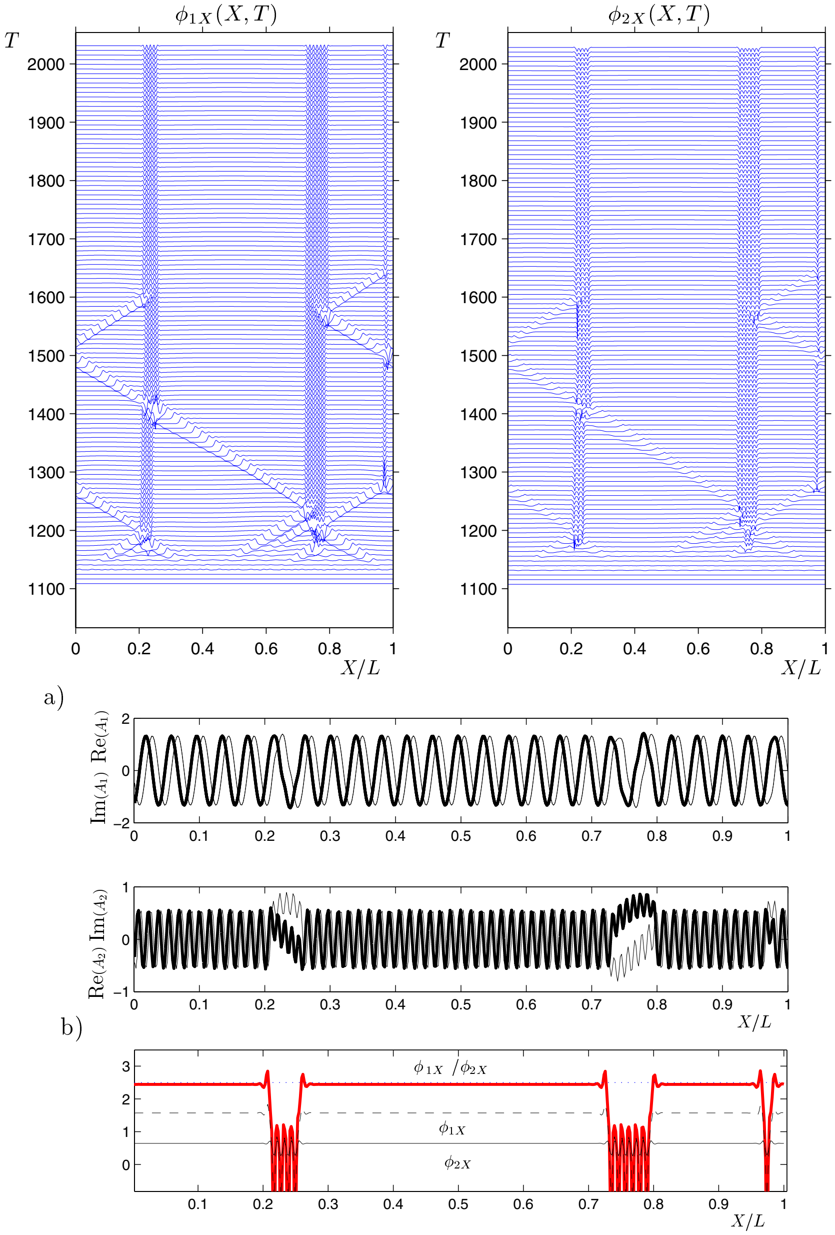


Figure 19. (a) *Space-time diagram of the phase-gradients amplitudes* A_1 *and* A_2*, respectively. Parameters as in Figure* 6(b) *but* $\gamma = 1$ *and with* $k = 0.5$*,* $\mu = 1.42$*, and* $L = 250$*. Lower figures correspond to the final state:* (a) *Real and imaginary part of the amplitudes* A_1 *and* A_2*;* (b) *phase gradients of* A_1 *and* A_2*. Dotted line indicates perfect* $2:5$ *resonance. The temporal evolution of* A_2 *is also shown in the corresponding movie* (60055_03.avi)*; red and yellow lines show the real and imaginary part of* A_2*, respectively, white its magnitude* $|A|$*, and green the local wavenumber. Clicking on the above image displays the associated movie.*

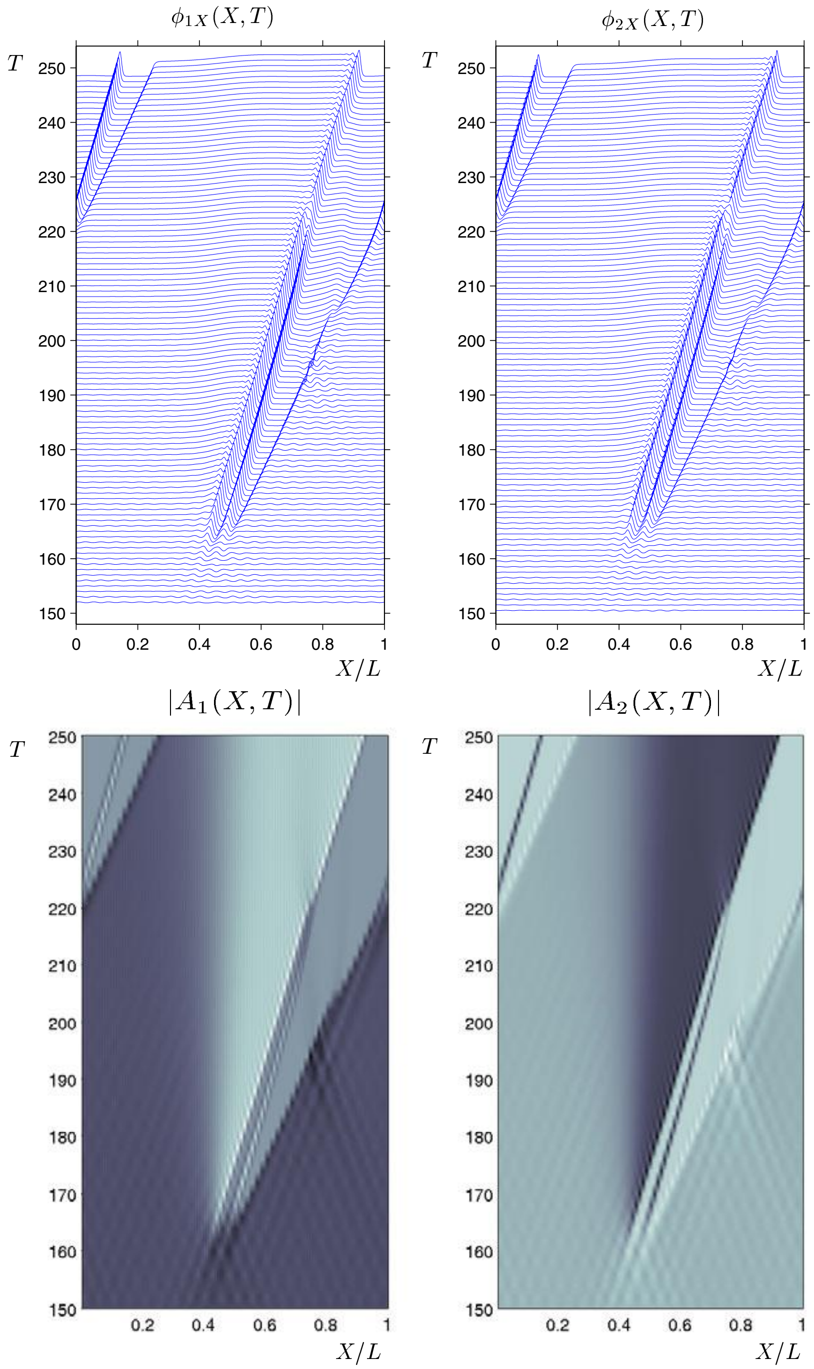


Figure 20. *Space-time diagram for the magnitude and the phase-gradient of the amplitudes A_1 and A_2. Parameters as in Figure* 6(b) *but $\gamma = 1.3$ and with $k = 0.5$, $\mu = 1.42$, and $L = 250$. In the gray-scale plots dark (light) stands for low (high) amplitudes. The temporal evolution of A_2 is also shown in the corresponding movie* (60055_04.avi)*; red and yellow lines show the real and imaginary part of A_2, respectively, white its magnitude $|A|$, and green the local wavenumber. Clicking on the above image displays the associated movie.*

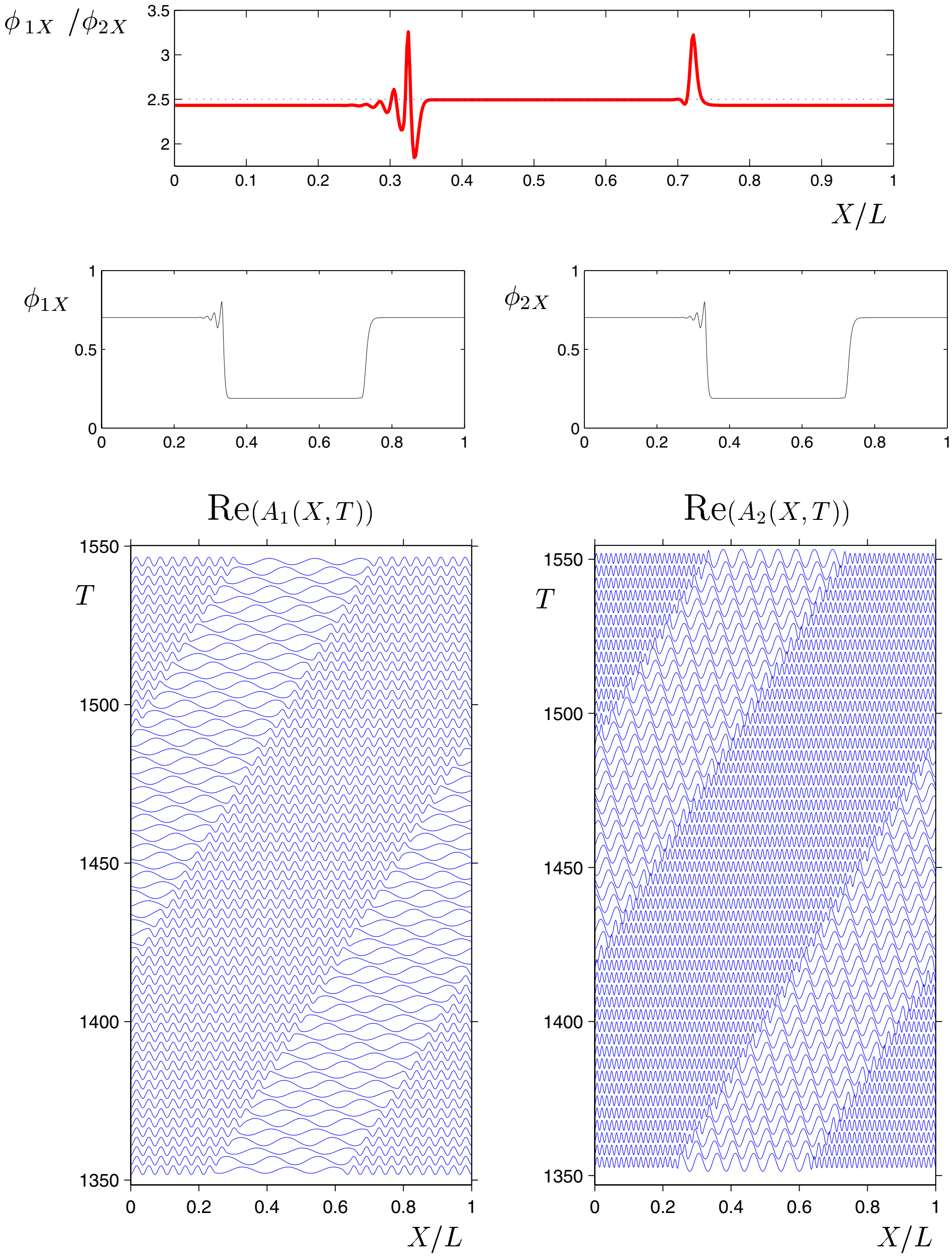


Figure 21. *Final, localized drift wave reached after the transient shown in Figure 20. Dotted line in plot of ϕ_{1X}/ϕ_{2X} indicates perfect $2:5$ resonance.*

stationary structures. Since the localized stationary structures often absorb the localized drift waves, the final state tends to be stationary in this case. If the initial perturbations are much smaller, fewer drift waves arise. If they travel in the same direction, no stationary localized structures are generated and the localized drift waves persist. Such a case is illustrated in Figures 20 and 21 and movie 60055_04.avi. Here, uniformly distributed perturbations with maximal amplitude 0.001 were applied independently to the real and imaginary part of both amplitudes. In this case only a single domain forms. It develops a defect in its interior, which eventually disappears at its trailing end (cf. movie 60055_04.avi). Initially, the propagation velocities of the leading and the trailing front of the domain are not equal, and the domain grows. As time goes on, however, the front velocities converge to a common value and one sees a stable, localized, propagating domain of fixed size containing traveling waves (see Figure 21). It is interesting to note that the wavenumbers selected in these traveling domains are in 2 : 5-resonance, while the stationary regions have slightly shifted wavenumbers that are not in resonance (see upper panel of Figure 21). The physical solution (cf. (2.2)) can therefore be described as consisting of regions of localized periodic waves traveling through a stationary quasi-periodic pattern. This type of solution is generally found when the side-band oscillatory instability occurs very close to the parity-breaking bifurcation. To describe the parity-breaking bifurcation in more detail the appropriate amplitude-phase equations [9, 43, 6, 4] could be derived from the coupled Ginzburg–Landau equations. They would, in particular, allow a better understanding of the localized drift waves. Such an undertaking is, however, beyond the scope of this paper.

Finally, in order to determine the character of the bifurcations we have also integrated (2.4), (2.5) using the final states as initial conditions and varying μ so as to take the system back into the stable regions of mixed modes. The results obtained indicate that the instabilities discussed in this section are subcritical (except for the cases shown in Figure 14). The solutions persist over a large interval of the forcing giving rise to bistability between periodic and quasi-periodic states. This behavior has been found in all cases we have considered. Therefore we expect that the coexistence of periodic and quasi-periodic stable states may be regarded as a characteristic property of the system (2.4), (2.5).

6. Conclusions.

In this paper we have investigated a near-resonant steady $m : n$ mode interaction in one-dimensional dissipative systems with reflection and translation symmetry using coupled Ginzburg–Landau equations that describe the slow evolution of the two relevant mode envelopes. One of the goals was to shed some light on the competition between periodic and quasi-periodic patterns by addressing in detail the stability of the periodic solutions, emphasizing particular instabilities that can lead to quasi-periodic patterns.

The simplest stationary solutions of this system are reflection-symmetric periodic patterns: the pure modes $S_{1,2}$ and the mixed modes $S_{\pm}$. Building on the results of Dangelmayr [13], we established the stability properties of these solutions with respect to perturbations that preserve the periodicity of the solution. Among the many possible choices of parameters we restricted our attention to sets of values that produced two supercritical primary bifurcations to pure modes followed by secondary bifurcations to mixed modes; these mixed modes, in turn, suffered symmetry-breaking bifurcations to traveling waves.

Side-band instabilities modify the stability regions of the pure and mixed modes. In particular, we find that these steady periodic patterns can undergo longwave as well as shortwave instabilities. For the pure modes the side-band instabilities are always of steady-state type. The destabilizing longwave perturbations correspond to the well-known Eckhaus instability and depend only on the linear coefficients of (2.4), (2.5). In contrast, the shortwave instabilities depend on nonlinear interactions and on the resonance considered (i.e., on m and n). In particular, we find the following.

1. If $m > 2$ and a pure mode, S_1, say, loses stability in a shortwave steady-state bifurcation, the wavenumber of the growing perturbation lies at the band center of the other mode (i.e., of S_2); this instability is affected by the detuning parameter $\hat{\gamma} = \gamma/2\delta$, which measures the mismatch between the two critical wavenumbers.

2. For wavenumber ratios of the form $2 : n$, one of the resonant terms enters into the stability calculation for the pure mode S_1. Away from the band center, this pure mode can experience shortwave instabilities, while close to the band center, it can lose stability only to homogeneous or longwave perturbations. In the shortwave case, due to the resonant terms the wavenumber selected by the instability is shifted with respect to the band center of the other pure mode, S_2. As the band center is approached the perturbation wavenumber of the shortwave instability can go to 0.

Close to their bifurcation from the pure modes, the mixed modes inherit the stability properties of the pure modes with respect to shortwave perturbations. In addition, they may undergo an oscillatory instability. It results from the interaction of the translation and the symmetry-breaking modes and can be strongly affected by the wavenumber mismatch.

The nonlinear evolution ensuing from the instabilities was investigated through direct numerical simulation of (2.4), (2.5) with periodic boundary conditions for the slowly varying envelopes. It was found that the transition from the pure modes to the mixed mode can depend strongly on the wavenumber mismatch. We have identified cases in which without a wavenumber mismatch the instability of the pure mode leads to a periodic mixed mode, while in the nonresonant case the instability of the pure mode is determined by a side-band instability resulting in a quasi-periodic pattern.

Side-band instabilities of the mixed modes generically destroy the spatial periodicity of the stationary, reflection-symmetric states and produce quasi-periodic patterns. They are typically characterized by a spatial modulation of the local wavenumber with a wavelength that is determined by the detuning parameter (in a manner predicted by the linear analysis) and the resonant terms (as indicated in (5.1)). The size of the modulation depends strongly on the strength of the resonance terms. For the $2 : 5$ case discussed here, as well as for resonances of the form $2 : n$ with $n > 5$, the effect is only noticeable for S_- (cf. Figure 17) because with S_+ the resonant terms are too small (cf. Figure 16). In contrast, the resonance terms in (2.4), (2.5) are comparable for $m : n = 2 : 3$ and their influence should be felt by both mixed modes.

The oscillatory instability tends to lead to the formation of propagating fronts separating reflection symmetric patterns from drift waves with broken reflection symmetry. The fronts may interact to form stable localized drift waves. These may subsequently collide with each other, destroying the initial periodic pattern completely in some cases and only partially in others. In the latter case the system relaxes to a very interesting pattern composed of alternating localized stationary domains of periodic and quasi-periodic states. We find other cases where all the fronts travel in the same direction, eventually achieving the same velocity and producing stable localized propagating domains of drift waves. The wavenumbers in these traveling domains are selected by the resonance condition (i.e., they are in the ratio m/n), whereas the wavenumbers associated with the surrounding steady periodic pattern are no longer resonant. Thus, the resulting patterns can be described as localized periodic traveling wave structures propagating through a stationary quasi-periodic background. Similar types of solutions have been found in the $1:2$ resonance [4] and may be expected in the general $m:n$ case whenever the resonant coefficients ν and ν' are of opposite sign.

In this paper we have focused on the periodic solutions and their stability. Complementary to this analysis would be a study of the quasi-periodic solutions. Of course, the set of solutions in which the two modes have nonresonant wavenumbers is considerably larger than that of the periodic solutions and such an analysis is beyond the scope of the present work. To understand how spatially quasi-periodic solutions arise from periodic ones, the analogy of the appearance of temporally quasi-periodic solutions through a Neimark–Sacker bifurcation might prove useful [27]. However, this framework, applied to the time-independent solutions, would not yield information about the stability of the quasi-periodic solutions.

In preliminary numerical simulations we have considered the stability of certain quasi-periodic solutions and found that the range in the wavenumber mismatch over which they are stable is smaller than that of periodic solutions with nearby wavenumbers [23]. Thus, at least in the regime considered, the resonant terms enhance the stability of the periodic solutions. Another promising line of investigation would be the reduction of the coupled Ginzburg–Landau equations (2.4), (2.5) to two coupled phase equations. This should be possible for longwave perturbations of the mixed-mode solution if the smallness of the resonant terms is exploited and it is assumed that such terms are of the same order as the gradients in the magnitude of the amplitudes (cf. (2.6), (2.7)).

Appendix A. Stability of pure and mixed modes.

In this appendix we provide some of the details of the stability analysis of the pure and mixed modes, presented in section 4.

A.1. Pure modes.

The perturbed pure mode S_1 is written as

$$A_1 = R_1 e^{ikX}\left(1 + a_1^+(T)e^{iQX} + a_1^-(T)e^{-iQX}\right), \tag{A.1}$$

$$A_2 = e^{ik\frac{n}{m}X}\left(a_2^+(T)e^{iQX} + a_2^-(T)e^{-iQX}\right). \tag{A.2}$$

Inserting this ansatz in (2.4), (2.5) and linearizing in $a_1^\pm$ and $a_2^\pm$, we obtain

$$\dot{a}_1^\pm = \left(\alpha - sR_1^2 - \delta Q^2 \mp (\gamma + 2\delta k)Q\right)a_1^\pm - sR_1^2 a_1^\mp, \tag{A.3}$$

$$\dot{a}_2^\pm = \left(\beta - \rho' R_1^2 - \delta' Q^2 \mp 2\delta' Q(nk/m)\right)a_2^\pm + \left\{\nu' R_1^n a_2^\mp\right\}_{m=2}, \tag{A.4}$$

where the bracketed term with subscript $m = 2$ is present only if $m = 2$, and α and β are defined by (3.3). The eigenvalues associated with (A.3) are

$$\lambda_1^\pm = -(\alpha + \delta Q^2) \pm \sqrt{\alpha^2 + Q^2(\gamma + 2k\delta)^2}, \tag{A.5}$$

and those associated with (A.4) are

$$\lambda_2^\pm = (\beta - \rho'\frac{\alpha}{s} - \delta' Q^2) \pm \sqrt{4\delta'^2\left(k\frac{n}{m}\right)^2 Q^2 + \left\{\nu'^2\left(\frac{\alpha}{s}\right)\right\}_{m=2}}. \tag{A.6}$$

Equations (A.5) and (A.6) are related to longwave and shortwave instabilities, respectively. Note that for $m \geq 3$ the perturbation amplitudes a_2^+ and a_2^- decouple and the destabilizing mode is composed of just one wave of the form $(a_2^+, 0)$ or $(0, a_2^-)$. In contrast, due to the resonant term when $m = 2$, the eigenvectors have (in principle) components in both directions and the destabilizing mode is composed of two waves with different wavenumbers.

For the pure mode S_2 we have

$$A_1 = e^{ikX}\left(a_1^+(T)e^{iQX} + a_1^-(T)e^{-iQX}\right), \tag{A.7}$$

$$A_2 = R_2\, e^{ik\frac{n}{m}X}\left(1 + a_2^+(T)e^{iQX} + a_2^-(T)e^{-iQX}\right), \tag{A.8}$$

where $a_1^\pm$ and $a_2^\pm$ satisfy

$$\dot{a}_1^\pm = \big(\alpha - \rho R_2^2 - \delta Q^2 \mp (\gamma + 2\delta k)Q\big)\, a_1^\pm, \tag{A.9}$$

$$\dot{a}_2^\pm = \big(\beta - 2s' R_2^2 - \delta' Q^2 \mp 2\delta'(nk/m)Q\big)\, a_2^\pm - s' R_2^2 a_2^\mp. \tag{A.10}$$

The stability of S_2 is determined by two eigenvalues:

$$\lambda_1^\pm = \left(\alpha - \rho\frac{\beta}{s'} - \delta Q^2\right) \pm |Q||\gamma + 2k\delta|,$$

$$\lambda_2^\pm = -(\beta + \delta' Q^2) \pm \sqrt{\beta^2 + 4\delta'^2\left(\frac{n}{m}k\right)^2 Q^2}, \tag{A.11}$$

which are associated with the shortwave and longwave instabilities, respectively.

A.2. Mixed modes. In this case we consider the system (2.6)–(2.9) for the evolution of the real amplitudes $R_j > 0$ and phases ϕ_j (for $j = 1, 2$). Thus (abusing notation) we write

$$R_1(X,T) = R_1 + r_1(X,T), \quad \phi_1(X,T) = kX + \varphi_1(X,T), \tag{A.12}$$

$$R_2(X,T) = R_2 + r_2(X,T), \quad \phi_2(X,T) = \frac{n}{m}kX + \varphi_2(X,T), \tag{A.13}$$

where

$$r_j = \left[r_j^+(T) - ir_j^-(T)\right] e^{iQX} + \left[r_j^+(T) + ir_j^-(T)\right] e^{-iQX},$$

$$\varphi_j = \left[\varphi_j^+(T) - i\varphi_j^-(T)\right] e^{iQX} + \left[\varphi_j^+(T) + i\varphi_j^-(T)\right] e^{-iQX}$$

for $j = 1, 2$. The linearized problem for the perturbations $r_j^\pm$ and $\varphi_j^\pm$ is given by

$$\begin{pmatrix} \dot{r}_1^\pm \\ \dot{\varphi}_1^\mp \\ \dot{r}_2^\pm \\ \dot{\varphi}_2^\mp \end{pmatrix} = \begin{pmatrix} a_1 & -2\delta Q(k+\gamma) & a_2 & 0 \\ -2\delta Q(k+\gamma) & b_1 & 0 & b_2 \\ c_2 & 0 & c_1 & -2\delta'\frac{nk}{m} \\ 0 & d_2 & -2\delta'\frac{nk}{m} & d_1 \end{pmatrix} \begin{pmatrix} r_1^\pm \\ \varphi_1^\mp \\ r_2^\pm \\ \varphi_2^\mp \end{pmatrix}, \tag{A.14}$$

where

$$a_1 = \mu - \delta k^2 - \gamma k - \delta Q^2 - (3sR_1^2 + \rho R_2^2) + \nu(n-1)R_1^{n-2}R_2^m \cos(n\hat{\phi}_1 - m\hat{\phi}_2),$$

$$a_2 = -2\rho R_1 R_2 + \nu m R_1^{n-1}R_2^{m-1},$$

$$b_1 = -\nu n R_1^{n-2}R_2^m - \delta Q^2, \quad b_2 = \nu m R_1^{n-1}R_2^{m-1},$$

$$c_1 = \mu - \Delta\mu - \delta'(kn/m)^2 - \delta' Q^2 - (3s'R_2^2 + \rho' R_1^2) + \nu'(m-1)R_2^{m-2}R_1^n \cos(n\hat{\phi}_1 - m\hat{\phi}_2),$$

$$c_2 = -2\rho' R_1 R_2 + \nu' n R_1^{n-1}R_2^{m-1},$$

$$d_1 = -\nu' m R_1^n R_2^{m-2} - \delta' Q^2, \quad d_2 = \nu' n R_1^{n-1}R_2^{m-1}.$$

Appendix B. Derivation of (4.9).

To derive (4.9) we first consider the characteristic equation in the form

$$\lambda^4 + G_3\lambda^3 + G_2\lambda^2 + G_1\lambda + G_0 = 0, \tag{B.1}$$

where G_j $(j = 0, \dots, 3)$ are polynomials in Q^2 whose coefficients depend on the parameters directly and also through the amplitudes of the mixed modes. When the bifurcation is steady and of shortwave type the location of Γ_+ and the wavenumber $Q \neq 0$ are obtained as solutions of the system $G_0 = \partial G_0/\partial Q^2 = 0$, which yields

$$(g_1 g_2)^2 - 4g_0 g_2^3 - 4g_1^3 - 18 g_0 g_1 g_2 = 0 \tag{B.2}$$

for the location of the lower part of Γ_+ and

$$Q^2 = (g_1 g_2 - 9g_0)/(6g_1 - 2g_2^2) \tag{B.3}$$

for the perturbation wavenumber of the most unstable perturbation. Here the functions g_0, g_1, and g_2 can be written as

$$\begin{aligned}
g_0 &= -(a_1c_1 - a_2c_2)(\delta' b_1 - \delta d_1) - \gamma^2 c_1 d_1,\\
g_1 &= \delta\delta'(a_1c_1 - a_2c_2 + a_1d_1 + c_1b_1) + \delta^2 c_1 d_1 + \delta'^2 a_1 b_1 + \gamma^2\delta'(c_1 + d_1),\\
g_2 &= -\delta^2\delta'(c_1 + d_1) - \delta'^2\delta(a_1 + b_1) - \gamma^2\delta'^2,
\end{aligned}$$

where a_j, b_j, c_j, and d_j $(j = 1, 2)$ are as defined in Appendix A.

To proceed, we consider the case $k = 0$ and make use of the fact that for $\gamma = 0$ the stability limit Γ^+ coincides with L_2 and remains close for $\gamma \ll 1$. We therefore expand the value of μ on the stability limit,

$$\mu = \mu_c + \eta^2, \quad \text{with} \quad \eta = c_1\gamma + c_2\gamma^2 + \cdots, \quad |\gamma| \ll 1, \tag{B.4}$$

where $\mu_c = \rho\Delta\mu/(s' - \rho)$ is the value of μ on the curve L_2 at $k = 0$. The small parameter η is a function of γ and must vanish when $\gamma = 0$. Moreover, the amplitudes of the mixed mode S_+ on Γ_+ may be expressed as

$$R_1 = \eta r_{11} + \eta^2 r_{12} + \cdots , \qquad R_2 = r_{20} + \eta^2 r_{22} + \eta^3 r_{23} \ldots , \tag{B.5}$$

where $r_{20} = \sqrt{\Delta\mu/(s' - \rho)}$ is the amplitude of the pure mode on the line L_2. The coefficients r_{ij} are determined by substituting the expansions (B.5) into (3.5) and solving at successive orders. The number of coefficients r_{ij} that must be kept depends on the particular resonance. In fact, the expansion (B.5) must be carried out to $O(\eta^{n-2})$ for an $m : n$ resonance. In order to calculate the c_j we substitute (B.5) into (B.2) and solve order by order. Finally, the wavenumber Q is obtained by substituting these results into (B.3).

We now sketch the necessary calculations for the case $m : n = 2 : 5$ at $k = 0$. Inserting (B.4) and (B.5) into (3.5) yields

$$\left.\begin{aligned} s r_{11}^2 + 2\rho r_{20} r_{22} = 1, \\ 2s' r_{20} r_{22} + \rho' r_{11}^2 = 1, \end{aligned}\right\} \quad \Longrightarrow \quad \begin{aligned} &r_{11}^2 = (s' - \rho)/(ss' - \rho\rho'), \\ &r_{22} = (s - \rho')/2r_{20}(ss' - \rho\rho') \end{aligned} \tag{B.6}$$

at $O(\eta^2)$, while at $O(\eta^3)$ we have

$$\left.\begin{aligned} 2s r_{11} r_{12} + 2\rho r_{20} r_{23} = -\nu r_{20}^m r_{11}^3, \\ s' r_{20} r_{23} + \rho' r_{11} r_{12} = 0, \end{aligned}\right\} \quad \Longrightarrow \quad \begin{aligned} &r_{12} = -\nu s' r_{11}^2 r_{20}^m / 2(ss' - \rho\rho'), \\ &r_{23} = \nu\rho' r_{11}^3 r_{20}^{m-1} / 2(ss' - \rho\rho'). \end{aligned} \tag{B.7}$$

If $n > 5$, we would find $r_{12} = r_{23} = 0$, and it would be necessary to consider higher-order corrections. Near the curve L_2 (i.e., near the point $(\mu, k) = (\mu_c, 0)$) (see Figure 6) the functions g_j, for $j = 0, 1, 2$, may be written as

$$\begin{aligned} g_0 &= \eta^n \left(4\delta'\nu(ss' - \rho\rho') r_{11}^n r_{20}^{m+2}\right) + O(\eta^{n+1}), \\ g_1 &= \eta^2 \left(4\delta\delta'(ss' - \rho\rho') r_{11}^2 r_{20}^2\right) + \eta^{n-2}(2s'n + \Delta\mu(n-1) r_{20}^m)\delta\delta'\nu a_1^{n-2} \\ &\quad + \gamma^2\delta'(-2s' r_{20}^2 - \eta^2 4s' r_{20} r_{22} - \eta^2 4s' r_{20} r_{23}) + O(\eta^n + \gamma^2\eta^{n-2}) \\ g_2 &= \delta^2\delta'(2s' r_{20}^2) + O(\eta^2), \end{aligned} \tag{B.8}$$

where only those terms necessary to compute Q have been kept in (B.8). Finally, composing the expansions (B.8) and (B.4) and substituting the result into (B.3), we find (after some algebra) the result (4.9).

Acknowledgments. We gratefully acknowledge discussions with J. Porter and P. Umbanhowar.

REFERENCES

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

[2] H. Arbell and J. Fineberg, *Pattern formation in two-frequency forced parametric waves*, Phys. Rev. E (3), 65 (2002), 036224.

[3] D. Armbruster, J. Guckenheimer, and P. Holmes, *Heteroclinic cycles and modulated travelling waves in systems with* O(2) *symmetry*, Phys. D, 29 (1988), pp. 257–282.

[4] A. Bayliss, B. Matkowsky, and H. Riecke, *Structure and dynamics of modulated traveling waves in cellular flames*, Phys. D, 74 (1994), pp. 1–23.

[5] T. Besson, W. Edwards, and L. Tuckerman, *Two-frequency parametric excitation of surface waves*, Phys. Rev. E (3), 54 (1996), pp. 507–513.

[6] B. Caroli, C. Caroli, and S. Fauve, *On the phenomenology of tilted domains in lamellar eutectic growth*, J. Phys. I (Paris), 2 (1992), pp. 281–290.

[7] P. Coullet, *Commensurate-incommensurate transition in nonequilibrium system*, Phys. Rev. Lett., 56 (1986), pp. 724–727.

[8] P. Coullet, C. Elphick, and D. Repaux, *Nature of spatial chaos*, Phys. Rev. Lett., 58 (1987), pp. 431–434.

[9] P. Coullet, R. Goldstein, and G. Gunaratne, *Parity-breaking transitions of modulated patterns in hydrodynamic systems*, Phys. Rev. Lett., 63 (1989), pp. 1954–1957.

[10] P. Coullet and P. Huerre, *Resonance and phase solitons in spatially forced thermal convection*, Phys. D, 23 (1986), pp. 27–44.

[11] J. D. Crawford, E. Knobloch, and H. Riecke, *Period-doubling mode interactions with circular symmetry*, Phys. D, 44 (1990), pp. 340–396.

[12] E. M. Curado and C. Elphick, *Effects of an almost resonant spatial thermal modulation in the Rayleigh-Bénard problem—quasi-periodic behavior*, J. Phys. A, 20 (1987), pp. 1205–1214.

[13] G. Dangelmayr, *Steady-state mode interactions in the presence of* O(2)*-symmetry*, Dyn. Stab. Syst., 1 (1986), pp. 159–185.

[14] J. H. Dawes, C. M. Postlethwaite, and M. R. E. Proctor, *Instabilities induced by a weak breaking of a strong spatial resonance*, Phys. D, 191 (2004), pp. 1–30.

[15] M. Dominguez-Lerma, D. Cannell, and G. Ahlers, *Eckhaus boundary and wave-number selection in rotating Couette-Taylor flow*, Phys. Rev. A, 34 (1986), p. 4956.

[16] A. Doelman and V. Rottschäfer, *Singularly perturbed and nonlocal modulation equations for sytems with interacting instability mechanisms*, J. Nonlinear Sci. 7 (1997), pp. 371–409.

[17] W. S. Edwards and S. Fauve, *Patterns and quasi-patterns in the Faraday experiment*, J. Fluid Mech., 278 (1994), pp. 123–148.

[18] G. Faivre, S. de Cheveigne, C. Guthmann, and P. Kurowski, *Solitary tilt waves in thin lamellar eutectics*, Europhys. Lett., 9 (1989), pp. 779–784.

[19] S. Fauve, S. Douady, and O. Thual, *Comment on "Parity-breaking transitions of modulated patterns in hydrodynamic systems,"* Phys. Rev. Lett., 65 (1990), p. 385.

[20] Z. C. Feng and P. R. Sethna, *Symmetry-breaking bifurcations in resonant surface waves*, J. Fluid Mech., 199 (1989), pp. 495–518.

[21] L. Fourtune, W. J. Rappel, and M. Rabaud, *Phase dynamics near a parity-breaking instability*, Phys. Rev. E, 49 (1994), pp. R3576–R3579.

[22] R. Goldstein, G. Gunaratne, L. Gil, and P. Coullet, *Hydrodynamic and interfacial patterns with broken space-time symmetry*, Phys Rev. A, 43 (1991), pp. 6700–6721.

[23] M. Higuera, H. Riecke, and M. Silber, unpublished results.

[24] M. Higuera, J. M. Vega, and E. Knobloch, *Coupled amplitude-streaming flow equations for nearly inviscid Faraday waves in small aspect ratio containers*, J. Nonlinear Sci., 12 (2002), pp. 505–551.

[25] C. A. Jones and M. R. E. Proctor, *Strong spatial resonance and travelling waves in Bénard convection*, Phys. Lett. A, 121 (1987), pp. 224–228.

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

[27] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Appl. Math. Sci. 112, Springer-Verlag, New York, 1995.

[28] H. Levine, W.-J. Rappel, and H. Riecke, *Resonant interactions and traveling solidification cells*, Phys. Rev. A, 43 (1991), pp. 1122–1125.

[29] M. Lowe, J. P. Gollub, and T. C. Lubensky, *Commensurate and incommensurate structures in a nonequilibrium system*, Phys. Rev. Lett., 51 (1983), pp. 786–789.

[30] B. A. Malomed and M. I. Tribelsky, *Bifurcations in distributed kinetic systems with aperiodic instability*, Phys. D, 14 (1984), pp. 67–87.

[31] E. Meron and I. Procaccia, *Low-dimensional chaos in surface waves: Theoretical analysis of an experiment*, Phys. Rev. A, 34 (1986), pp. 3221–3237.

[32] P. Metzener and M. R. E. Proctor, *Interaction of patterns with disparate scales*, Eur. J. Mech. B Fluids, 11 (1992), pp. 759–778.

[33] J. Moehlis and E. Knobloch, *Forced symmetry breaking as a mechanism for bursting*, Phys. Rev. Lett., 80 (1998), pp. 5329–5332.

[34] A. Ogawa, W. Zimmermann, K. Kawasaki, and T. Kawakatsu, *Forced periodic and quasi-periodic patterns in anisotropic systems*, J. Phys. II (Paris), 6 (1996), pp. 305–328.

[35] E. Pampaloni, P. L. Ramazza, S. Residori, and F. T. Arecchi, *Two-dimensional crystals and quasi-crystals in nonlinear optics*, Phys. Rev. Lett., 74 (1995), pp. 258–261.

[36] J. Porter and E. Knobloch, *Complex dynamics in the* $1:3$ *spatial resonance*, Phys. D, 143 (2000), pp. 138–168.

[37] J. Porter and E. Knobloch, *New type of complex dynamics in the* $1:2$ *spatial resonance*, Phys. D, 159 (2001), pp. 125–154.

[38] J. Porter and M. Silber, *Broken symmetries and pattern formation in two-frequency forced Faraday waves*, Phys. Rev. Lett., 89 (2002), 084501.

[39] M. R. E. Proctor and C. A. Jones, *The interaction of two spatially resonant patterns in thermal convection. Part* 1. *Exact* $1:2$ *resonance*, J. Fluid Mech., 188 (1988), pp. 301–335.

[40] M. Rabaud, S. Michalland, and Y. Couder, *Dynamical regimes of directional viscous fingering: Spatiotemporal chaos and wave propagation*, Phys. Rev. Lett., 64 (1990), pp. 184–187.

[41] W.-J. Rappel and H. Riecke, *Parity-breaking in directional solidification: Numerics versus amplitude equations*, Phys. Rev. A, 45 (1992), pp. 846–859.

[42] H. Riecke and H. G. Paap, *Stability and wave-vector restriction of axisymmetric Taylor vortex flow*, Phys. Rev. A, 33 (1986), pp. 547–559.

[43] H. Riecke and H.-G. Paap, *Parity-breaking and Hopf bifurcation in axisymmetric Taylor vortex flow*, Phys. Rev. A, 45 (1992), pp. 8605–8610.

[44] J. Rogers, M. Schatz, O. Brausch, and W. Pesch, *Superlattice patterns in vertically oscillated Rayleigh-Bénard convection*, Phys. Rev. Lett., 85 (2000), pp. 4281–4284.

[45] A. M. Rucklidge and W. J. Rucklidge, *Convergence properties of the* $8, 10,$ *and* 12 *mode representations of quasipatterns*, Phys. D, 178 (2003), pp. 62–82.

[46] A. Simon, J. Bechhoefer, and A. Libchaber, *Solitary modes and the Eckhaus instability in directional solidification*, Phys. Rev. Lett., 61 (1988), pp. 2574–2577.

[47] F. Simonelli and J. P. Gollub, *Surface-wave mode interactions: Effects of symmetry and degeneracy*, J. Fluid Mech., 199 (1989), pp. 471–494.

[48] M. Umeki, *Faraday resonance in rectangular geometry*, J. Fluid Mech., 227 (1991), pp. 161–192.

[49] R. Wiener and D. McAlister, *Parity-breaking and solitary waves in axisymmetric Taylor vortex flow*, Phys. Rev. Lett., 69 (1992), pp. 2915–2918.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 4, pp. 503–524

A Vaccination Model for Transmission Dynamics of Influenza[*]

M. E. Alexander[†], C. Bowman[‡], S. M. Moghadas[†], R. Summers[†], A. B. Gumel[§], and B. M. Sahai[¶]

Abstract. Despite the availability of preventive vaccines and public health vaccination programs, influenza inflicts substantial morbidity, mortality, and socio-economic costs and remains a major public health problem. This is largely because the protection conferred by current vaccines is dependent on the immune status of the individual, ranging between 70%–90% in healthy young adults and 30%–40% among the elderly and others with weakened immune systems. Whether a strategic use of such partially effective vaccines can control the spread of influenza within a certain population is unknown but of great public health interest. To address this question, we construct a deterministic mathematical model to study the transmission dynamics of influenza. The model is analyzed qualitatively to determine criteria for control of an influenza epidemic and is used to compute the threshold vaccination rate necessary for community-wide control of influenza. Using two specific populations of similar size, an office and a personal care home, our model shows that the spread of influenza can be controlled if the combined effect of the vaccine efficacy and vaccination rate reaches a threshold determined by the duration of infectiousness and the rate of contact between infected and susceptible individuals.

Key words. epidemic models, bistability, influenza, outbreaks, vaccination

AMS subject classifications. Primary, 34C23, 92B05; Secondary, 37N25

DOI. 10.1137/030600370

1. Introduction.

Influenza (also known as flu) is a respiratory disease caused by certain RNA viruses of the *Orthomyxoviridae* family [13, 22, 30]. These viruses have a segmented RNA genome that is replicated by an error-prone RNA polymerase leading to seasonal mutations (drifts) in parts of the viral genome. Human influenza viruses appear as three distinct serotypes: A, B, and C. Only influenza A viruses infect and multiply in avian species and nonprimate mammals, and their genomic segments are subject to occasional reassortment (shift) giving rise to new viral strains consisting of novel hemagglutinin (HA) and/or neuraminidase (NA) genes [23, 30, 39] (see Figure 1). The influenza A viruses have been responsible for the vast majority of epidemics and all recorded pandemics (see Table 1), which usually result from

[*] Received by the editors June 18, 2003; accepted for publication (in revised form) by J. Yorke July 13, 2004; published electronically November 9, 2004. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

http://www.siam.org/journals/siads/3-4/60037.html

[†] Institute for Biodiagnostics, National Research Council Canada, Winnipeg, Manitoba, Canada R3B 1Y6 (murray.alexander@nrc-cnrc.gc.ca, seyed.moghadas@nrc-cnrc.gc.ca, randy.summers@ nrc-cnrc.gc.ca).

[‡] Institute for Biodiagnostics, National Research Council Canada, Winnipeg, Manitoba, Canada R3B 1Y6 (christopher.bowman@nrc-cnrc.gc.ca); Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.

[§] Department of Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2 (gumelab@cc.umanitoba.ca).

[¶] Cadham Provincial Public Health Laboratory, Winnipeg, Manitoba, Canada R3C 3Y1 (bsahai@gov.mb.ca).

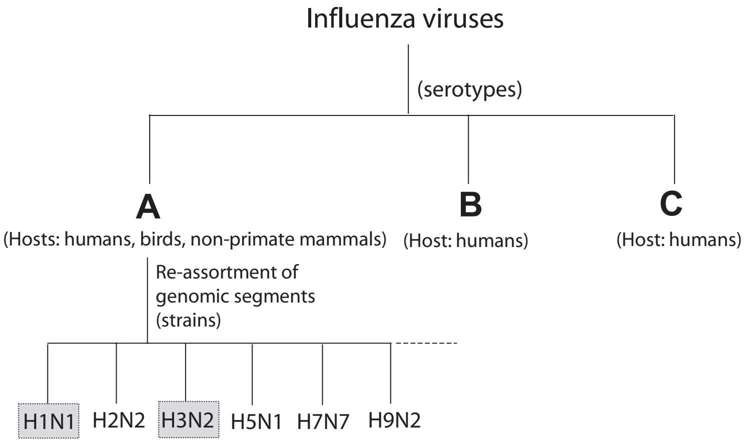

Influenza viruses
(serotypes)
A
(Hosts: humans, birds, non-primate mammals)
B
(Host: humans)
C
(Host: humans)
Re-assortment of genomic segments (strains)
H1N1
H2N2
H3N2
H5N1
H7N7
H9N2

Figure 1. *Types and known strains of human influenza viruses. Influenza A viruses are primarily associated with epidemics in humans, and sometimes their replication in a host coinfected with more than one viral strain leads to new viral strains due to the reassortment of HA and NA subtype genes. The individual strains further diverge into independent lineages by genetic drift. The $H1N1$ and $H3N2$ are the most common strains circulating in the human population* [39].

Table 1

Influenza viral strains and outbreaks and epidemics. Most influenza outbreaks and epidemics, including all pandemics of the last century, have been caused by the spread of a single influenza virus strain consisting of a specific HA and NA subtypes. This is regardless of the cocirculation of any other viral strain during that influenza season.

Year	Outbreak/epidemic region	Viral strain (lineage)	Refs
1918	Pandemic (Spanish flu)	H1N1 (A/Brevig Mission/1/18; A/South Carolina/1/18; and A/New York/1/18)	[31, 39]
1957	Pandemic (Asian flu)	H2N2 (A/Singapore/1/57)	[31, 39]
1968	Pandemic (Hong Kong flu)	H3N2 (A/NT/60/68; and A/Hong Kong/1/68)	[31, 39]
1977	Pandemic in children and young adults (Russian flu)	H1N1 (A/USSR/90/77)	[31, 39]
2001-02	USA, Canada, Singapore, Malaysia, Egypt, Europe	H1N2 (A/New Caledonia/20/99 H1 and A/Moscow/10/99 N2)	[41]
2001-02	Northern Italy	Influenza B (B/Victoria/2/87)	[2]
2001-02	St. Elisabeth Hospital, Tilburg	H3N2 (A/Sydney/5/97)	[6]
2003	Poultry farms in Netherlands	H7N7 (Avian influenza A)	[21]
2003-04	USA, Canada, Europe, Japan	H3N2 (A/Fujian/411/2002)	[37, 38]

uncontrolled replication and spread of a single virus strain [25, 39]. While concurrent circulation of two or more viral strains during an influenza season is not uncommon, an outbreak in a localized population usually involves a single strain [1]. Due to drift in the viral genome, a lasting protective immunity against influenza viruses has been an elusive target to achieve.

Influenza viruses have coexisted with humans for centuries and have historically been a cause of excessive morbidity and mortality [9, 10, 34]. Annually, the virus affects 25 to 50 million people, with an estimated 20 to 40 thousand influenza-related deaths, in the United States [34]. Because of the illness and high number of deaths associated with influenza, particularly among the elderly [10, 11], much attention has been focused on preventive strategies

[13, 17, 25]. Although vaccination has been an effective strategy against influenza infection [5, 12, 17, 26, 28, 34, 40], current preventive vaccines consisting of inactivated virions do not protect all vaccine recipients equally. The vaccine-based protection is dependent on the immune status of the recipient (see [18, 35] for general references). Typically, influenza vaccines protect 70%–90% of the recipients among healthy young adults and as low as 30%–40% of the elderly and others with weakened immune systems (such as HIV-infected or immunosuppressed transplant patients) [5, 12, 15, 26]. Furthermore, due to the seasonal drift in the viral genome, annual vaccination against the influenza virus strains anticipated to be in circulation during the upcoming season is necessary to prevent new infections and subsequent outbreaks.

The failure of current influenza vaccines to protect all vaccine recipients warrants the determination of conditions necessary for a substantial reduction, approaching eradication, of influenza infection in a population. Consequently, the aim of this study is to explore, via mathematical modeling, the impact of immunization with a partially effective vaccine on the transmission dynamics of influenza infection. The study addresses the question of whether such a vaccine could ever completely stop the spread of infection and determines the minimal vaccine efficacy and vaccination rate required to control or eradicate infection in a population.

Mathematical models have been used to determine the ability of an imperfect vaccine to control other infectious diseases, and some of the findings have been corroborated by clinical studies (see [14, 16, 19, 20, 27] for general references). There have been several published mathematical models suggested for the transmission dynamics of influenza [4, 13, 18, 23, 36], but to our knowledge none has fully analyzed the impact of an imperfect vaccine (see also [24] and the references therein). Furthermore, these studies tend to classify all recruited individuals into the population as susceptibles. In the context of influenza epidemiology, it is more realistic to consider models that also allow for the continuous recruitment of infected individuals into the population, as previously considered for other diseases such as HIV infection [7]. In the model presented here, the use of an imperfect vaccine and recruitment of infected individuals prove to be crucial factors determining the transmission dynamics of influenza.

The paper is organized as follows. A mathematical model for the transmission dynamics of influenza infection is formulated and normalized in section 2. The existence of the equilibria of the normalized-reduced (NR) model is discussed in section 3. Stability analysis of the NR model is carried out in section 4, where it is also shown that the model has a stable endemic equilibrium even when the threshold condition for disease eradication holds. In section 5, some quantitative results are derived from the model to illustrate the effect of vaccination in two typical environments consisting of equal-size populations: a personal care home and an office setting. We conclude with a brief discussion of our findings.

2. Model formulation.

In order to derive the equations of the mathematical model, we divide the population ($\tilde{N}$) into four subpopulations: susceptible ($\tilde{S}$), vaccinated ($\tilde{V}$), infected ($\tilde{I}$), and recovered ($\tilde{R}$). Since a typical outbreak of influenza is caused by the replication and spread of a single viral strain [1], irrespective of the existing immunity to previous strains, our model monitors the dynamics of influenza based on a single strain without effective cross-immunity against the strain. It should be noted, however, that the model does not exclude the possibility of two concurrent outbreaks, each caused by a different strain. In this case, the

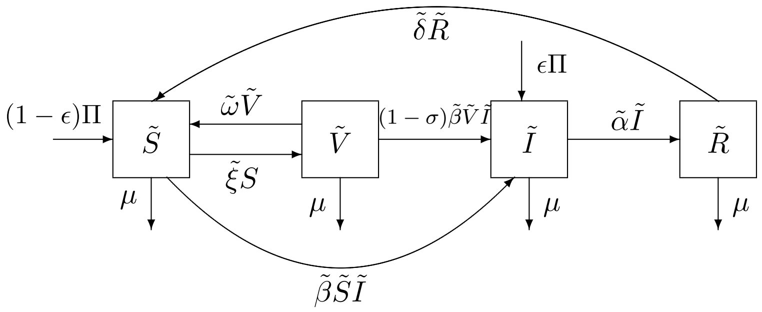

$\tilde{\delta}\tilde{R}$
$\epsilon\Pi$
$(1-\epsilon)\Pi$
$\tilde{\omega}\tilde{V}$
$(1-\sigma)\tilde{\beta}\tilde{V}\tilde{I}$
$\tilde{\alpha}\tilde{I}$
$\tilde{S}$
$\tilde{V}$
$\tilde{I}$
$\tilde{R}$
$\tilde{\xi}S$
μ
μ
μ
μ
$\tilde{\beta}\tilde{S}\tilde{I}$

Figure 2. *Transfer diagram of the* $\tilde{S}\tilde{V}\tilde{I}\tilde{R}\tilde{S}$ *model.*

model does not consider the effects of partial cross-immunity between the viral strains.

The susceptible population is increased by recruitment of individuals (either by birth or immigration), and by the loss of immunity, acquired through previous vaccination or natural infection. This population is reduced through vaccination (moving to class $\tilde{V}$) and infection (moving to class $\tilde{I}$), and by natural death or emigration.

The population of vaccinated individuals is increased by vaccination of susceptibles. Since the vaccine does not confer immunity to all vaccine recipients, vaccinated individuals may become infected but at a lower rate than unvaccinated (those in class $\tilde{S}$). The vaccinated class is thus diminished by this infection (moving to class $\tilde{I}$) and further decreased by waning of vaccine-based immunity (moving to class $\tilde{S}$) and by natural death.

The population of infected individuals is increased by recruitment of infected individuals from outside the population, as well as by infection of susceptibles including those who remain susceptible despite being vaccinated. It is diminished by natural death and by recovery from the disease (moving to class $\tilde{R}$).

Since the immunity acquired by infection wanes with time, the recovered individuals become susceptible to the disease again. Thus the recovered class is increased by individuals recovering from their infection and is decreased as the natural immunity wanes (moving back to class $\tilde{S}$).

The transfer diagram for these processes is shown in Figure 2. The details of the transitions between the subpopulations can be mathematically expressed by the following differential equations:

$$\frac{d\tilde{S}}{d\tilde{t}} = (1-\epsilon)\Pi + \tilde{\omega}\tilde{V} + \tilde{\delta}\tilde{R} - \tilde{\beta}\tilde{S}\tilde{I} - \tilde{\xi}\tilde{S} - \mu\tilde{S}, \tag{2.1}$$

$$\frac{d\tilde{V}}{d\tilde{t}} = \tilde{\xi}\tilde{S} - (1-\sigma)\tilde{\beta}\tilde{V}\tilde{I} - (\tilde{\omega}+\mu)\tilde{V}, \tag{2.2}$$

$$\frac{d\tilde{I}}{d\tilde{t}} = \epsilon\Pi + \tilde{\beta}\tilde{S}\tilde{I} + (1-\sigma)\tilde{\beta}\tilde{V}\tilde{I} - (\tilde{\alpha}+\mu)\tilde{I}, \tag{2.3}$$

$$\frac{d\tilde{R}}{d\tilde{t}} = \tilde{\alpha}\tilde{I} - (\mu+\tilde{\delta})\tilde{R}, \tag{2.4}$$

where Π is the rate of recruitment of individuals into the population; ϵ is the fraction of recruited individuals who are already infected; $\tilde{\xi}$ is the rate at which susceptible individuals

Table 2
Model parameters and their interpretations.

Parameter	Description
Π	recruitment rate of individuals
ϵ	fraction of recruited individuals who are infected
$\tilde{\beta}$	$\frac{\text{contacts}}{\text{time}}\times$probability of infection per contact with an infected
$\tilde{\xi}$	rate at which susceptible individuals are vaccinated
σ	vaccine efficacy
$1/\tilde{\omega}$	average time to lose vaccine-induced immunity
$1/\tilde{\alpha}$	average length of infection (duration of infectiousness)
$1/\tilde{\delta}$	average time to lose immunity acquired by infection
$1/\mu$	average life-span

receive the vaccine; μ is the rate at which people leave the population, whether by death or emigration (this rate is assumed to be the same for all subpopulations); $\tilde{\beta}$ represents the probability of infection for susceptible individuals; $\tilde{\omega}$ is the rate at which vaccine-based immunity wanes; σ is the vaccine efficacy; $\tilde{\alpha}$ is the recovery rate from infection; and $\tilde{\delta}$ is the rate of loss of immunity acquired by infection.

It is worth noting that recruitment of individuals into the vaccinated class ($\tilde{V}$) does not significantly alter the dynamics of the model (see the appendix). Furthermore, since the model monitors the dynamics of the human populations, it is assumed that all the model parameters and state variables are nonnegative. A description of the model parameters is given in Table 2.

To simplify the mathematical analysis of this study, we normalize the model (2.1)–(2.4) by defining the new variables,

$$S = \frac{\mu}{\Pi}\tilde{S}, \quad V = \frac{\mu}{\Pi}\tilde{V}, \quad I = \frac{\mu}{\Pi}\tilde{I}, \quad R = \frac{\mu}{\Pi}\tilde{R},$$

and parameters,

$$t = \mu\tilde{t}, \quad \beta = \frac{\Pi\tilde{\beta}}{\mu^2}, \quad \omega = \frac{\tilde{\omega}}{\mu}, \quad \xi = \frac{\tilde{\xi}}{\mu}, \quad \delta = \frac{\tilde{\delta}}{\mu}, \quad \alpha = \frac{\tilde{\alpha}}{\mu},$$

giving

$$\frac{dS}{dt} = (1-\epsilon) - \beta SI - \xi S - S + \omega V + \delta R, \tag{2.5}$$

$$\frac{dV}{dt} = \xi S - (1-\sigma)\beta VI - (1+\omega)V, \tag{2.6}$$

$$\frac{dI}{dt} = \epsilon + \beta SI + (1-\sigma)\beta VI - (1+\alpha)I, \tag{2.7}$$

$$\frac{dR}{d\tilde{t}} = \alpha I - (1+\delta)R. \tag{2.8}$$

Let $N = S+V+I+R$ be the total population size of the model (2.5)–(2.8). Adding equations (2.5)–(2.8) gives the equation for the total population:

$$\frac{dN}{dt} = 1 - N. \tag{2.9}$$

Since $N(t) \to 1$ as $t \to \infty$, it can be seen that the feasible region

$$\Omega = \{(S, V, I, R): \ S, V, I, R \geq 0; \ S + V + I + R = 1\}$$

is positively invariant for the model (2.5)–(2.8). Therefore, we restrict our attention to the dynamics of the model in Ω. Using $R = 1 - S - V - I$ in Ω, (2.8) can be removed from the model. This substitution gives the following NR model:

$$\frac{dS}{dt} = (1-\epsilon) - \beta SI - \xi S - S + \omega V + \delta(1 - S - V - I), \tag{2.10}$$

$$\frac{dV}{dt} = \xi S - (1-\sigma)\beta VI - (1+\omega)V, \tag{2.11}$$

$$\frac{dI}{dt} = \epsilon + \beta SI + (1-\sigma)\beta VI - (1+\alpha)I. \tag{2.12}$$

3. Equilibria of the NR model. In this section, the conditions for the existence of the equilibria of the NR model, given by (2.10)–(2.12), are established. In order to do this, we first consider the case $\epsilon = 0$; that is, the population does not admit new infected individuals. The results obtained in this case will subsequently be adapted and used in section 3.3 to discuss the existence of the equilibria of the NR model with $\epsilon > 0$.

3.1. Disease-free equilibrium (DFE). When $\epsilon = 0$, the model assumes that all individuals recruited into the population are susceptible. In this case, the NR model has a DFE given by

$$E_0 = \left(\frac{1+\omega}{1+\omega+\xi}, \frac{\xi}{1+\omega+\xi}, 0 \right).$$

To establish the conditions for the existence of endemic equilibria when $\epsilon = 0$, it is useful to analyze the stability of E_0. This analysis provides a key threshold quantity which will be used for stability analysis of the NR model throughout the paper. To determine the local stability of E_0, the Jacobian of the NR model is evaluated at the DFE to yield

$$J_0 = \begin{pmatrix} -(1+\delta+\xi) & \omega - \delta & -\dfrac{\beta(1+\omega)}{1+\omega+\xi} - \delta \\ \xi & -(1+\omega) & -\dfrac{(1-\sigma)\beta\xi}{1+\omega+\xi} \\ 0 & 0 & \dfrac{\beta(1+\omega)}{1+\omega+\xi} + \dfrac{(1-\sigma)\beta\xi}{1+\omega+\xi} - (1+\alpha) \end{pmatrix}.$$

The eigenvalues of J_0 are

$$\lambda_1 = \frac{\beta(1+\omega)}{1+\omega+\xi} + \frac{(1-\sigma)\beta\xi}{1+\omega+\xi} - (1+\alpha),$$

and the eigenvalues of the submatrix

$$J_1 = \begin{pmatrix} -(1+\delta+\xi) & \omega - \delta \\ \xi & -(1+\omega) \end{pmatrix}.$$

It is easy to see that the eigenvalues of J_1 are negative. Thus, the local stability of E_0 depends on the sign of λ_1. Let

$$\mathscr{R}_0 = \frac{\beta[(1+\omega)+(1-\sigma)\xi]}{(1+\alpha)(1+\omega+\xi)}. \tag{3.1}$$

Lemma 3.1. *The DFE of the NR model is locally asymptotically stable if $\mathscr{R}_0 < 1$ and unstable if $\mathscr{R}_0 > 1$.*

Proof. Since the eigenvalues of J_1 are negative, it follows that the DFE is locally asymptotically stable if $\lambda_1 < 0$ and unstable if $\lambda_1 > 0$. Noting that $\lambda_1 < 0$ if and only if $\mathscr{R}_0 < 1$, the proof is complete. ■

The threshold quantity $\mathscr{R}_0$ in (3.1) is the reproductive number of infection [3] which can be interpreted as the number of infected people produced by one infected individual introduced into the population in the presence of vaccination (see [35]). Biologically speaking, Lemma 3.1 implies that the disease can be eradicated from the population (when $\mathscr{R}_0 < 1$) if the initial sizes of the subpopulations are in the basin of attraction of E_0. This is, however, an insufficient condition for disease control, since for *arbitrary* initial sizes of the subpopulations, the local stability of E_0 does not guarantee community-wide eradication of the disease. This scenario will be discussed in section 4.

The quantity $\mathscr{R}_0$ can be rewritten as

$$\mathscr{R}_0 = \left(1 - \frac{\sigma\xi}{1+\omega+\xi}\right) r_0, \tag{3.2}$$

where $r_0 \equiv \frac{\beta}{1+\alpha}$ is the basic reproductive number of the disease in the absence of vaccination [3], in which case $\mathscr{R}_0$ reduces to r_0. Clearly, if $r_0 < 1$, then $\mathscr{R}_0 < 1$ irrespective of the value of ξ, and thus E_0 is locally asymptotically stable.

Using the expression for $\mathscr{R}_0$ in (3.2) and the fact that $\xi \geq 0$, it can be seen that

$$(1-\sigma)r_0 = \frac{\beta}{\beta^*} < \mathscr{R}_0 \leq r_0,$$

where

$$\beta^* \equiv \frac{1+\alpha}{1-\sigma}.$$

Thus, if $\beta > \beta^*$, then $\mathscr{R}_0 > 1$, and consequently no amount of vaccination can bring $\mathscr{R}_0$ below 1. The requirement $\beta < \beta^*$ can be rewritten so as to define a threshold vaccine efficacy, σ_c:

$$\sigma_c = 1 - \frac{1}{r_0}. \tag{3.3}$$

If $\sigma < \sigma_c$, then $\beta > \beta^*$, and $\mathscr{R}_0 > 1$ regardless of vaccination rate. Figure 3 shows the regions of vaccine efficacy and vaccination rate for which $\mathscr{R}_0 < 1$ for some chosen parameter values.

From now on, it is assumed that β satisfies

$$1+\alpha < \beta < \frac{1+\alpha}{1-\sigma} \equiv \beta^*. \tag{3.4}$$

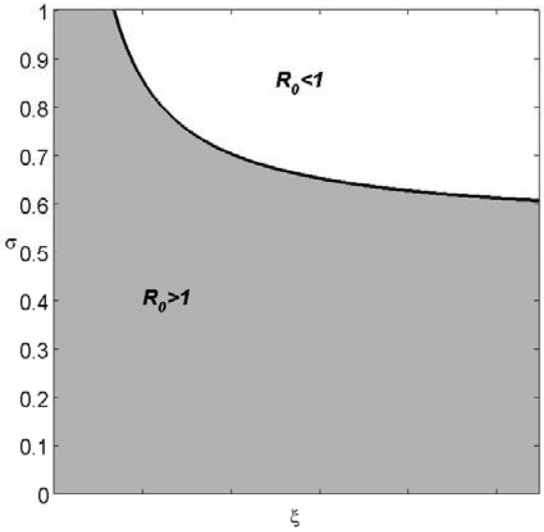


Figure 3. *Regions for reducing basic reproduction number to less than unity given an imperfect vaccine with efficacy σ applied at rate ξ. The gray and white regions correspond to $\mathscr{R}_0 > 1$ and $\mathscr{R}_0 < 1$, respectively. Other model parameters are $\alpha = 4500$, $\delta = 1000$, $\omega = 100$, $\sigma = 0.9$, and $\beta = 10000$.*

In this case, setting $\mathscr{R}_0 = 1$ and solving for ξ give the threshold vaccination rate

$$\xi_c = \frac{(1+\omega)(r_0-1)}{1-(1-\sigma)r_0}, \tag{3.5}$$

which is positive.

3.2. Endemic equilibria ($\epsilon = 0$). The endemic equilibria of the NR model with $\epsilon = 0$ (if they exist) cannot be cleanly expressed in closed form. In order to find the conditions for the existence of these equilibria, we use (2.10)–(2.12) to express the variables S and V in terms of the variable I when $I \neq 0$. This gives (at equilibrium)

$$S = \frac{(1-\sigma)\beta I + 1 + \omega}{[(1-\sigma)(\beta I + \xi) + 1 + \omega]r_0}, \tag{3.6}$$

$$V = \frac{1 - r_0 S}{(1-\sigma)r_0}. \tag{3.7}$$

Substituting (3.6)–(3.7) into (2.12) gives (at equilibrium)

$$Q(I) \equiv a_1 I^2 + a_2 I + a_3 = 0, \tag{3.8}$$

where

$$a_1 = \beta^2(1-\sigma)(1+\alpha+\delta), \tag{3.9}$$

$$a_2 = \beta\{(1+\alpha+\delta)[(1-\sigma)\xi + 1 + \omega] - (1+\delta)(1-\sigma)[\beta - (1+\alpha)]\}, \tag{3.10}$$

$$a_3 = (1+\delta)\{(1+\alpha)(1+\omega+\xi) - \beta[(1-\sigma)\xi + 1 + \omega]\}. \tag{3.11}$$

Since all the model parameters are nonnegative, it follows from (3.9) that $a_1 > 0$. Furthermore, if $\mathscr{R}_0 > 1$, then $a_3 < 0$. The existence of the equilibria is summarized in the following theorem and illustrated in Figure 5.

Theorem 3.2. *Suppose $\epsilon = 0$ in (2.10)–(2.12).*

(a) *If $a_2(\xi_c) \geq 0$, then*

(i) *the model has a unique endemic equilibrium if $\xi < \xi_c$,*

(ii) *the model has no endemic equilibria if $\xi > \xi_c$.*

(b) *If $a_2(\xi_c) < 0$, then there exists $\xi^* > \xi_c$ such that*

(i) *the model has a unique endemic equilibrium if $\xi < \xi_c$,*

(ii) *the model has no endemic equilibria if $\xi > \xi^*$,*

(iii) *the model has two endemic equilibria if $\xi_c < \xi < \xi^*$.*

Proof. Since $\mathscr{R}_0$ is a continuous decreasing function of ξ for $\xi > 0$, if $\xi < \xi_c$, then $\mathscr{R}_0 > 1$ ($a_3 < 0$). Since $a_1 > 0$, it follows that $Q(I)$ has a unique positive root.

Suppose $\xi \geq \xi_c$. In this case, $\mathscr{R}_0 \leq 1$ ($a_3 > 0$). It is easy to see that $a_2(\xi)$ is an increasing function of ξ. Thus, if $a_2(\xi_c) \geq 0$, then $a_2(\xi) > 0$ for $\xi > \xi_c$, and $Q(I)$ has no positive real root, which implies that the model has no endemic equilibrium in this case. If $a_2(\xi_c) < 0$, consider $D(\xi) = a_2^2(\xi) - 4a_1a_3(\xi)$, the discriminant of $Q(I)$. Since $a_3(\xi_c) = 0$, $D(\xi_c) > 0$, and $D(\xi)$ is a quadratic function of ξ with positive coefficient for ξ^2. Furthermore, since $a_2(\xi)$ is a linear increasing function of ξ, there is a unique $\xi^{**} > \xi_c$ so that $a_2(\xi^{**}) = 0$, and thus $D(\xi^{**}) < 0$ (since a_1 and a_3 are both nonnegative for $\xi \geq \xi_c$). Let ξ^* be the unique root of $D(\xi)$ in $[\xi_c, \xi^{**}]$. Then, $a_2(\xi) < 0$, $a_1 > 0$, $a_3 \geq 0$, and $D(\xi) > 0$ for $\xi \in (\xi_c, \xi^*)$. Hence, $Q(I)$ has two positive roots, so that the model has two endemic equilibria, if $\xi_c < \xi < \xi^*$. Taking into account $a_2(\xi) > 0$ for $\xi > \xi^{**}$ and $D(\xi) < 0$ for $\xi \in (\xi^*, \xi^{**})$, it follows that the model has no endemic equilibria if $\xi > \xi^*$.

If $\mathscr{R}_0 = 1$ ($a_3 = 0$), then $Q(I)$ reduces to $(a_1 I + a_2)I = 0$. In this case, the NR model has a unique endemic equilibrium if $a_2 < 0$ and no endemic equilibrium if $a_2 \geq 0$. ∎

3.3. Equilibria ($\epsilon > 0$).

Since influenza can also be introduced into the population by the recruitment of infected individuals, it is more realistic to consider the NR model with $\epsilon > 0$. Mathematically speaking, if recruitment of infected individuals is allowed, the DFE does not exist, and eradication of the disease may not be feasible. In this case the public health objective is to minimize the level of epidemicity.

Consider now the NR model (2.10)–(2.12) with $\epsilon > 0$. It can be shown that the equilibria of the model are now the roots of the following cubic:

$$\begin{aligned} P(I,\epsilon) &= a_1 I^3 + a_2 I^2 + [a_3 - \epsilon\beta(1+\delta)(1-\sigma)]I - \epsilon(1+\delta)(\xi+\omega+1) \\ &= IQ(I) - \epsilon\left[\beta(1+\delta)(1-\sigma)I + (1+\delta)(\xi+\omega+1)\right], \end{aligned} \tag{3.12}$$

where $Q(I)$ is defined in (3.8).

Notice that when $\epsilon = 0$, $P(I)$ has roots which correspond to the roots of $Q(I)$ (along with $I = 0$). Furthermore, $P(I,\epsilon) < IQ(I)$ for $\epsilon > 0$ and $I > 0$. We now consider three cases as follows.

Case 1: $a_3 < 0$. In this case, $Q(I)$ has a unique positive root, denoted by I^*. Since $P(I,\epsilon)$ is a decreasing function of ϵ for positive I, it follows that $P(I^*,\epsilon) < 0$ for $\epsilon > 0$. Furthermore, $P(I,\epsilon) \to \infty$ as $I \to \infty$. Thus, $P(I,\epsilon)$ has a unique positive root for all $\epsilon \geq 0$, and this unique positive root must be at $I > I^*$. That is, when the NR model has a unique endemic equilibrium with $\epsilon = 0$, recruitment of infected individuals introduces no new equilibria but serves to shift the existing (unique) equilibrium to a higher disease state.

Case 2: $a_3 > 0$ and $a_2 > 0$. Let η_1, η_2, and η_3 represent the roots of $P(I, \epsilon)$. In this case, since $a_1 > 0$ and $a_2 > 0$, it follows that $\eta_1 + \eta_2 + \eta_3 = -a_2/a_1 < 0$ and $\eta_1\eta_2\eta_3 > 0$. This implies that $P(I)$ has a unique positive root for any $\epsilon > 0$.

Case 3: $a_3 > 0$ and $a_2 < 0$. If $a_2^2 - 4a_1a_3 > 0$, then from Theorem 3.2 it follows that $Q(I)$ has two positive roots. Thus, $P(I, 0)$ has two positive roots if $a_2^2 - 4a_1a_3 > 0$. Since $P(I, \epsilon)$ is a cubic function of I and is a decreasing function of ϵ for $I > 0$, it can be seen that there is a positive ϵ^* such that $P(I, \epsilon)$ has three positive roots if $0 < \epsilon < \epsilon^*$; two positive roots (one with multiplicity 2) if $\epsilon = \epsilon^*$; and a unique positive root if $\epsilon > \epsilon^*$. It should be noted that $\epsilon^* < \epsilon_0$, where

$$\epsilon_0 = \frac{a_3}{\beta(1-\sigma)(1+\delta)}.$$

To see this, suppose $\epsilon \geq \epsilon_0$. If $P(I, \epsilon)$ has exactly one real root, then it follows from $\eta_1\eta_2\eta_3 = \epsilon(1 + \delta)(\xi + \omega + 1)/a_1 > 0$ that this root must be positive. If the roots of $P(I, \epsilon)$ are all real, since

$$\eta_1\eta_2 + \eta_1\eta_3 + \eta_2\eta_3 = \frac{a_3 - \epsilon\beta(1+\delta)(1-\sigma)}{a_1} \leq 0,$$

when $\epsilon \geq \epsilon_0$, it can be seen that $P(I, \epsilon)$ also has a unique positive root in this case.

The above discussion shows that if the NR model has multiple endemic equilibria when $\epsilon = 0$, increasing ϵ to $\epsilon > \epsilon^*$ serves to remove this phenomenon and instead shifts the NR model to a higher epidemicity which is characterized by a unique endemic equilibrium with a higher number of infected individuals than for the case $\epsilon = 0$. These results are summarized below.

Theorem 3.3. (a) *If $\xi < \xi_c$ or $\epsilon > \epsilon_0$, then the model* (2.10)–(2.12) *has a unique positive endemic equilibrium.*

(b) *If $\xi > \xi_c$ and $a_2 > 0$, then the model* (2.10)–(2.12) *has a unique positive endemic equilibrium for $\epsilon > 0$.*

(c) *If $\xi > \xi_c$, $a_2 < 0$, and $a_2^2 - 4a_1a_3 > 0$, then there is a positive $\epsilon^* < \epsilon_0$ such that $P(I)$ has*

(i) *three positive roots if $0 < \epsilon < \epsilon^*$;*

(ii) *two positive roots if $\epsilon = \epsilon^*$;*

(iii) *a unique positive root if $\epsilon > \epsilon^*$.*

When $\xi > \xi_c$, $a_2 < 0$, and $a_2^2 - 4a_1a_3 < 0$, the existence of the equilibria of the model is more complicated. In this case, the model can have one, two, or even three endemic equilibria for $\epsilon > 0$. Bifurcation diagrams for a set of parameter values and various values of ϵ, depicted in Figure 6, support this claim. However, since $P(I)$ is a decreasing function of ϵ, $P(0, 0) = 0$ and $\lim_{I\to\infty} P(I, \epsilon) = +\infty$, it follows that $P(I, \epsilon > 0)$ has at least one positive root and hence, at least one endemic equilibrium always exists.

4. Stability analysis of the NR model.

In order to establish the stability of the equilibria of the NR model, we first investigate the nonexistence of certain types of solutions such as homoclinic orbits, periodic orbits, or polygons. From (2.10)–(2.12), it can be seen that the following region is positively invariant for the NR model:

$$\Omega_1 = \{(S, V, I) : \quad S, V, I \geq 0, \; S + V + (1 + d)I = 1\},$$

where $d \equiv \frac{\alpha}{1+\delta}$. Therefore, the *omega-limit* set of each solution of the NR model is contained in Ω_1. Thus, by applying Lemma 3.1 in [8] to the model equations (2.10)–(2.12), the following can be shown.

Lemma 4.1. *The NR model* (2.10)–(2.12) *has no periodic orbits, homoclinic orbits, or polygons (heteroclinic cycles) in* Ω_1.

4.1. Global stability with $\epsilon = 0$.

From Theorem 3.2, it is clear that if $\xi < \xi_c$, then the NR model has a unique endemic equilibrium (which is located in Ω_1) and the DFE is unstable. Since Ω_1 is positively invariant, it follows from Lemma 4.1 that the *omega-limit* set of each solution of the NR model in $\Omega_1 \setminus \Omega_0$ must be the endemic equilibrium of the model, where

$$\Omega_0 = \{(S, V, 0) \in \Omega_1 : \; S + V = 1\}.$$

It is easy to see that the DFE attracts the solutions in Ω_0. Thus, the unique endemic equilibrium of the NR model is globally asymptotically stable in $\Omega_1 \setminus \Omega_0$.

Suppose $\xi > \xi_c$. In this case, the DFE is the only equilibrium of the NR model if $a_2 > 0$ or $a_2^2 - 4a_1a_3 < 0$ (Theorem 3.2). Since Ω_1 is positively invariant, it follows from Lemma 4.1 that the DFE is globally asymptotically stable. Therefore, we have established the following theorem.

Theorem 4.2. (a) *If* $\xi < \xi_c$, *then the unique endemic equilibrium of the NR model is globally asymptotically stable in* $\Omega_1 \setminus \Omega_0$.

(b) *Suppose* $\xi > \xi_c$. *The DFE* (E_0) *is globally asymptotically stable if one of the following statements holds:*

(i) $a_2(\xi_c) \geq 0$;

(ii) $a_2(\xi_c) < 0$ *and* $\xi > \xi^*$.

Now consider the case where the NR model has two endemic equilibria ($\xi_c < \xi < \xi^*$). Using the expressions for ξ_c and r_0, the coefficient a_3 can be rewritten as

$$a_3 = (1 + \delta)(1 + \alpha)[1 - (1 - \sigma)r_0](\xi - \xi_c).$$

Noting that $1 < r_0 < 1/(1 - \sigma)$ (from (3.4)), it follows that $a_3 > 0$ as long as $\xi > \xi_c$. Otherwise, the model has a unique endemic equilibrium which is globally asymptotically stable (by Theorem 4.2).

Here, we shall show that the two endemic equilibria of the NR model cannot be repellers (in Ω_1) simultaneously. Thus, one of these equilibria must have a stable manifold in Ω_1. This stable manifold is located on a curve which is in the interior of Ω_1.

Lemma 4.3. *Suppose* $\xi_c < \xi < \xi^*$ *and* $a_2(\xi_c) < 0$. *Then the two endemic equilibria of the NR model cannot both be repellers in* Ω_1 *simultaneously. Furthermore, the DFE is not globally asymptotically stable.*

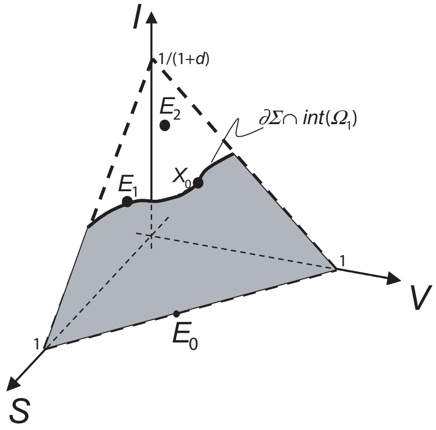


Figure 4. *Feasible region in SVI space including three equilibria E_0, E_1, and E_2. The gray region shows the basin of attraction of E_0 in Ω_1, while the white region shows the basin of attraction of E_2 in Ω_1. The curve $\partial\Sigma \cap \mathrm{int}(\Omega_1)$ is the stable manifold of E_1 which separates the two basins of attraction.*

Proof. Let Γ be the basin of attraction of E_0 and $\Sigma = \Gamma \cap \Omega_1$. Furthermore, since E_0 attracts Ω_0, it follows that $\Omega_0 \subset \Sigma$. Since Γ is an open set, it can be seen that $\partial\Sigma \cap \mathrm{int}(\Omega_1) \neq \varnothing$, where $\partial\Sigma$ is the boundary of Σ and $\mathrm{int}(\Omega_1)$ is defined to be $\Omega_1 \backslash \partial\Omega_1$. Suppose $X^0 = (S^0, V^0, I^0)$ is an arbitrary point in $\partial\Sigma \cap \mathrm{int}(\Omega_1)$, and let $\Phi(t, X^0)$ be a solution of the NR model (2.10)–(2.12) with $\Phi(0, X^0) = X^0$ (see Figure 4). Since the endemic equilibria are located in the interior of Ω_1, we can pick X^0 such that X^0 is different from the two endemic equilibria, denoted by E_1 and E_2. Since X^0 is not in the basin of attraction of E_0, it follows that $\Phi(t, X^0)$ cannot converge to E_0. Since $X^0 \in \Omega_1$ and Ω_1 is positively invariant, it follows that the *omega-limit* set of the solution $\Phi(t, X^0)$ must be in $\Omega_1 \setminus \Sigma$. Furthermore, since the model has no periodic orbits in Ω_1 (Lemma 4.1), the solution $\Phi(t, X^0)$ must converge to one of the two endemic equilibria. This implies that both endemic equilibria cannot be repellers in Ω_1 simultaneously. Thus, although E_0 is locally asymptotically stable (since $\mathscr{R}_0 < 1$), this equilibrium is not globally asymptotically stable. ■

The above result shows that one of the two endemic equilibria (namely, E_1) has at least a one-dimensional stable manifold. If E_1 is in the interior of $\Omega_1 \setminus \partial\Sigma \cap \mathrm{int}(\Omega_1)$, then since the basin of attraction of E_0 is an open set, a discontinuity appears in the direction field of the model in $\partial\Sigma \cap \mathrm{int}(\Omega_1)$. Note that solutions with initiating points very close to X^0, but in the basin of attraction of E_0, approach E_0. Hence, E_1 is located in $\partial\Sigma \cap \mathrm{int}(\Omega_1)$. Therefore, the stable manifold of E_1 separates Ω_1 into two basins of attraction. Consequently, the other endemic equilibrium (namely, E_2) is stable. In summary, in this case, the model has two stable equilibria and a saddle endemic equilibrium (E_1) where the stable manifold of E_1 separates the basins of attraction of E_0 and E_2. This is summarized in the following theorem.

Theorem 4.4. *Suppose $\xi_c < \xi < \xi^*$ and $a_2(\xi_c) < 0$. Then the NR model has bistable equilibria.*

The epidemiological implication of this theorem is that a vaccination program with $\xi > \xi_c$ ($\mathscr{R}_0 < 1$) would not guarantee disease eradication. In this case, the initial sizes of the subpopulations determine whether the disease can be eradicated from the population. If these subpopulations are initially in the basin of attraction of the locally stable endemic equilibrium (E_2), the disease will persist, and similarly, the disease can be eradicated if they are in the basin of attraction of the DFE (E_0).

4.2. Region of bistability with $\epsilon = 0$. Here, we shall show that under certain conditions, there is a positive interval such that the NR model with $\epsilon = 0$ undergoes the phenomenon of bistability (possessing a stable endemic equilibrium along with the stable DFE) for any β in this interval.

By Theorem 3.2, bistability is determined by the sign of $a_2(\xi_c)$, which is given by

$$a_2(\xi_c) = \frac{\beta(b_2\beta^2 + b_1\beta + b_0)}{(1+\alpha)[1-(1-\sigma)r_0]}, \tag{4.1}$$

where

$$b_0 = (1+\alpha)\left((1+\alpha+\delta)(1+\sigma\omega) + (1-\sigma)\alpha\delta\right), \tag{4.2}$$

$$b_1 = -(1-\sigma)(2-\sigma)(1+\alpha)(1+\delta), \tag{4.3}$$

$$b_2 = (1-\sigma)^2(1+\delta). \tag{4.4}$$

Since $(1+\alpha) < \beta < \beta^*$, it follows that $1-(1-\sigma)r_0 > 0$. Thus, the sign of $a_2(\xi_c)$ is determined by the sign of the quadratic $f(\beta) = b_2\beta^2 + b_1\beta + b_0$. Consider the discriminant $b_1^2 - 4b_0b_2$ of this quadratic, and let

$$\Delta = \frac{\sigma(1+\delta)(1+\alpha)}{4(1+\omega)(1+\alpha+\delta)}.$$

Noting that $b_1^2 - 4b_0b_2 > 0$ if and only if $\Delta > 1$ and that $b_0, b_2 > 0$ and $b_1 < 0$, it can be seen that $f(\beta)$ has two positive roots, β_1 and β_2 (with $\beta_1 < \beta_2$), if $\Delta > 1$. Then, bistability occurs when $\beta_1 < \beta < \beta_2$ and $\beta < \beta^*$. It remains to show that this intersection is nonempty. It is clear that $\beta_1 < \beta_m$, where $\beta_m = \frac{-b_1}{2b_2}$ is the value of β which minimizes $f(\beta)$. Thus,

$$\beta_1 < \beta_m = \frac{(2-\sigma)(1+\alpha)}{2(1-\sigma)} < \frac{1+\alpha}{1-\sigma} = \beta^*,$$

and hence, for $\beta_1 < \beta < \min(\beta_2, \beta^*)$, the NR model with $\epsilon = 0$ has two endemic equilibria for some value of ξ (see Figure 5). Therefore, we have established the following theorem.

Theorem 4.5. *If $\Delta > 1$, then there exists a positive interval (with nonzero measure) such that $a_2(\xi_c) < 0$ for any β in this interval. Furthermore, there is a range of $\xi > \xi_c$ such that the NR model (with $\epsilon = 0$) exhibits bistability.*

Proof. The proof trivially follows from Theorems 3.2 and 4.4. ■

It should be noted that

$$\Delta < \frac{1+\delta}{4(1+\omega)},$$

irrespective of the value of the parameter α, and thus if $\delta < 4\omega$, then $\Delta < 1$. Therefore, bistability is only possible in our model if the immunity acquired by infection (δ) wanes at least

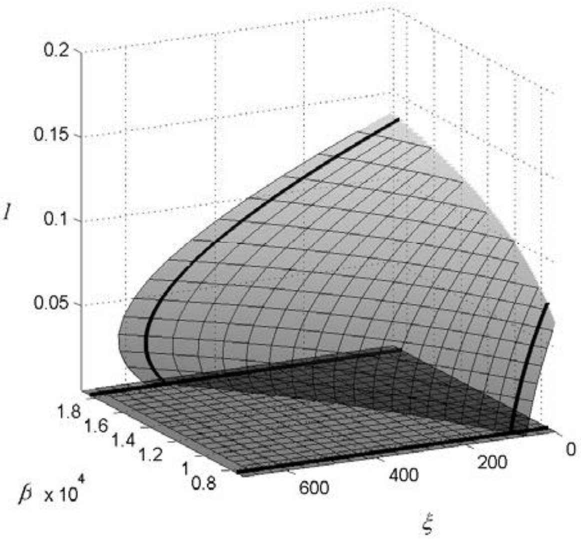


Figure 5. *Graph of the roots of quadratic $Q(I)$ as a function of the parameters ξ and β using $\alpha = 4500$, $\delta = 1000$, $\omega = 100$, and $\sigma = 0.9$. For $\mu = 0.02$ (year)$^{-1}$ (representing an average life-expectancy of 50 years), the values of α, δ, ω represent an infectious period of 4 days, an average period of 10 days for the loss of immunity acquired by infection, and a period of 6 months for the loss of immunity induced by vaccine, respectively. The solid lines show the bifurcation behavior at $\beta = 7000$ (transcritical bifurcation) and $\beta = 17000$ (backward bifurcation), respectively.*

four times faster than the vaccine-induced immunity (ω); this is, however, unlikely because the immunity acquired by natural infection is usually more robust and lasts longer than that induced by a vaccine consisting of inactivated virus particles such as current influenza vaccines [29]. Thus, this study has shown that the coexistence of multiple stable solutions is not feasible in influenza epidemics, which implies that the threshold condition $\mathscr{R}_0 < 1$ guarantees eradication of the disease within the population.

4.3. Stability analysis with $\epsilon > 0$. Since Ω_1 is a positively invariant region for the NR model, it follows that the equilibria of the model are contained in Ω_1. Thus, if the NR model has a unique endemic equilibrium, then it follows from Lemma 4.1 that this equilibrium is globally asymptotically stable. The epidemiological implication of this result is that the disease will persist within the population.

It is important to note that disease eradication is not biologically feasible as long as $\epsilon > 0$ (that is, the model allows for recruitment of infected individuals into the population). This fact can mathematically be shown by the following discussion. When $\epsilon > 0$, the model can have one, two, or even three endemic equilibria which all are contained in the positively invariant region Ω_1 (see Figure 4). Since the two-dimensional simplex Ω_1 is bounded and the model has no periodic orbits, homoclinic orbits, or polygons, it follows from the Poincaré–Bendixson theorem [33] that the *omega-limit* set of every solution in Ω_1 is an equilibrium. Note that the model has no DFE, so that the disease will persist within the population.

What is perhaps more important is that, although increasing the vaccination rate does not lead to disease eradication, it can prevent the high level of epidemicity which could occur for a large proportion of infected individuals. To verify this fact, we use the equation $P(I) = 0$ (at equilibrium) to obtain $\xi = G(I)/H(I)$ as a function of I, where

$$\begin{aligned} G(I) &= \beta^2(1-\sigma)(1+\alpha+\delta)I^3 + \beta[(1-\sigma)(1+\delta)(1+\alpha-\beta) + (1+\omega)(1+\alpha+\delta)]I^2 \\ &\quad + [(1+\delta)(1+\omega)(1+\alpha-\beta) - \epsilon\beta(1-\sigma)(1+\delta)]I - \epsilon(1+\omega)(1+\delta), \end{aligned}$$

and

$$H(I) = -\beta(1-\sigma)(1+\alpha+\delta)I^2 - (1+\delta)[1+\alpha-\beta(1-\sigma)]I + \epsilon(1+\delta).$$

Let I_0 be the unique positive root of $H(I)$. It is easy to show that $G(I)$ (and, consequently, $\xi(I)$) has a unique positive root. Let I_1 denote this root. We shall show that $I_1 > I_0$. It is clear that I_1 corresponds to an endemic equilibrium of the NR model when $\xi = 0$. Thus, it can be seen that I_1 is a root of the following quadratic:

$$E(I) = -\beta(1+\alpha+\delta)I^2 - (1+\delta)(1+\alpha-\beta)I + \epsilon(1+\delta).$$

A simple calculation yields

$$L(I) \equiv E(I) - H(I) = \beta\sigma I[1+\delta-(1+\alpha+\delta)I].$$

The quadratic $L(I)$ has two roots, namely, $I = 0$ and $I_* = \frac{1+\delta}{1+\alpha+\delta}$. It is easy to see that $E(I_*) < 0$. Since $L(I_*) = 0$, it follows that $H(I_*) < 0$ and thus, $I_* > I_0$. Noting that $L(I) > 0$ for $I \in (0, I_*)$, it can be seen that $I_1 > I_0$. This implies that $\xi(I) \geq 0$ for $I \in (I_0, I_1]$ and $\xi(I) < 0$ for $I \in [0, I_0) \cup \{I : \ I > I_1\}$. Since $\xi \to +\infty$ when $I \to I_0^+$, it follows that increasing vaccination coverage reduces the number of infected individuals at equilibrium, with I_0 representing the smallest possible level of endemicity.

The results of this section are summarized in the following theorem.

Theorem 4.6. *Suppose* $\epsilon > 0$.

(a) *If the NR model has a unique endemic equilibrium, then it is globally asymptotically stable. Furthermore, the NR model has no DFE and the disease will persist within the population.*

(b) *Increasing vaccination coverage reduces the number of infected individuals (at equilibrium) and prevents the high level of epidemicity.*

5. Outbreaks in two typical populations.

We now provide a discussion of two specific examples that illustrate the quantitative aspects of the model. We consider the spread of influenza in two representative populations: one consisting of a geriatric population in a personal care home, and the other healthy office workers. These populations were chosen in view of the fact that outbreaks of influenza frequently occur among the elderly in personal care homes who inadequately respond to the vaccine [11, 15, 28]. The dynamics of influenza in this population was compared to that seen amongst office workers who in general adequately respond to the vaccine.

We assumed the same steady state population $\Pi/\mu = 100$ in each case and set the natural death rate (including the rate of emigration) for the personal care home and office to $\mu =$

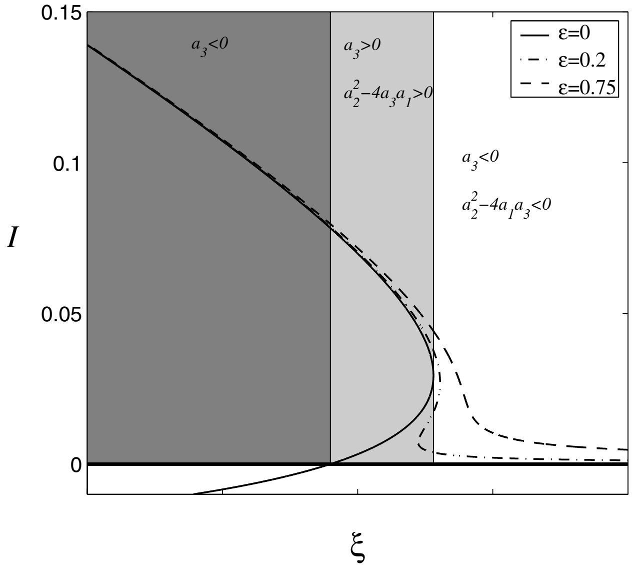


Figure 6. *Bifurcation diagram showing I at equilibrium versus vaccination rate for three values of ϵ, the fraction of recruited individuals who are infected. Other parameters are $\alpha = 4500$, $\delta = 1800$, $\omega = 100$, and $\sigma = 0.9$.*

Table 3
Parameter values for the study populations.

Parameter	Office	Personal care home	Unit
Steady state population, Π/μ	100	100	-
Death and emigration rate, μ	0.05	0.35	yr^{-1}
Effective waning rate of immunity, $\tilde{\delta}$	1	1	yr^{-1}
Waning rate of vaccine-induced immunity, $\tilde{\omega}$	1	1	yr^{-1}
Vaccine efficacy, σ	0.8	0.3	-
Transmission rate, $\tilde{\beta}$	4.5	4.5	$\frac{\text{yr}^{-1}}{\text{person}}$
Recovery period, $1/\tilde{\alpha}$	4	20	days
Vaccination rate, $\tilde{\xi}$	variable	variable	yr^{-1}

0.35 yr^{-1} and $\mu = 0.05$ yr^{-1}, respectively. The higher rate of μ for the personal care home accounts for excessive deaths common to such a population. As shown in Table 3, the two populations differ in their ability to elicit protective immunity in response to the vaccine (the vaccine efficacy, σ) [5, 12, 15, 26, 28] and their rate of recovery from influenza infection ($\tilde{\alpha}$) [32]. The effective waning rate of immunity refers to the failure of any prior infection-induced immunity against the circulating influenza virus strain(s). The values of rates of waning of effective immunity and vaccine-induced immunity both reflect the annual drift in the virus genome. The value for the rate of transmission ($\tilde{\beta}$) is taken within the range estimated by Nuño et al. [32]. The parameter values are summarized in Table 3, and bifurcation diagrams illustrating the relative fraction of infected cases as a function of vaccination rate in each population are shown in Figure 7.

The model clearly shows that, due to notably lower efficacies of influenza vaccines in

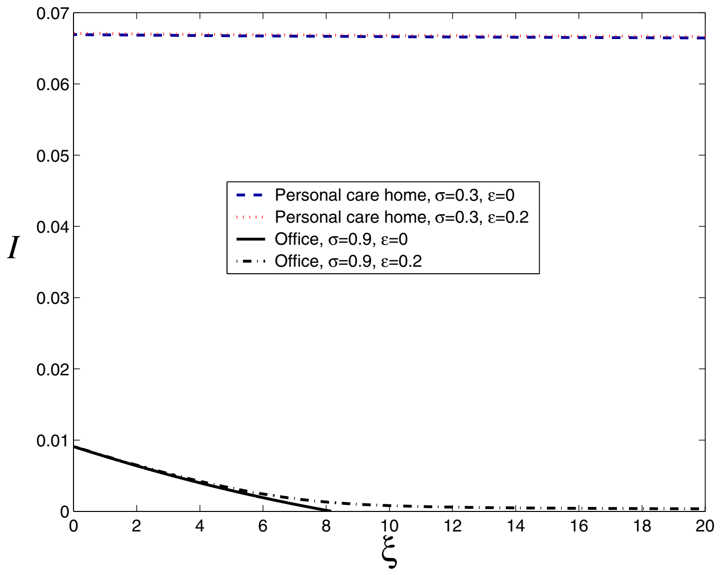


Figure 7. *Fraction of the infected population versus vaccination rate for the two study populations. The solid and dashed-dot curves represent the profiles of infected individuals in the office setting, where the vaccine efficacy is* 90%*, without (*$\epsilon = 0$*) and with (*$\epsilon = 0.2$*) the recruitment of infected individuals into the office population, respectively. The color dashed and dotted curves show the same profiles in a personal care home with a vaccine efficacy of* 30%*. Evidently, infection ratios in the personal care home are considerably higher, and vaccination has little effect on controlling the outbreak.*

personal care homes ($\sigma = 0.3$) [15, 28] than in the office setting ($\sigma = 0.8$), the former is more susceptible to a large-scale influenza outbreak. Vaccination would therefore have little effect in controlling outbreaks in personal care homes. In fact, the parameters in Table 3 for the personal care home yield $\sigma < \sigma_c$, and thus no amount of vaccination (with a partially effective vaccine) can control the outbreak. Likewise, reducing the rate at which infected individuals are recruited would have little effect in the personal care home scenario as well as in the sparsely vaccinated workers in the office setting. On the other hand, adding a continuous inflow of infected people to the office situation with high vaccination rate will drive the system away from its disease-free state into an endemic state with low infection. Simulation of the model using a larger population ($\Pi/\mu = 1000$) in both scenarios yields results that are qualitatively consistent with those reported in Figure 7, albeit in this case the epidemicity levels are higher.

We have chosen not to consider a large heterogeneous population (such as a city) for the present study, mainly on account of the lack of reliable data on demography which significantly affects the level of vaccine-based protection. Factors affecting the results of such a study include the proportion and distribution of immune-suppressed (such as HIV-infected and geriatric) individuals and the nature of interaction between susceptible and infected individuals. With specified demography and parameter values, our model will be able to illustrate the dynamics of influenza irrespective of the size of the population.

6. Conclusion. This paper evaluates the impact of a partially effective preventive vaccine on the control of influenza infection, using a new deterministic mathematical model. The model allows for continuous recruitment of individuals into the susceptible and infected populations. It should be noted that this model can be applied to other infectious diseases that also exhibit transitions between subpopulations as shown in Figure 2. In fact, this model extends the classical model for infectious diseases, such as measles and whooping cough, with $\delta = 0$; that is, the immunity acquired by infection is permanent (see [3] for a general reference).

In the absence of recruitment of infected individuals, linear stability analysis shows that the model has a DFE which is locally asymptotically stable when the basic reproductive number ($\mathscr{R}_0$) [3] is less than unity and unstable otherwise. However, in general, the local stability of the DFE does not necessarily imply its global stability, and the stable DFE and an endemic equilibrium may coexist. Several epidemic models have been proposed in which bistability has been observed, particularly in models in which the population is divided into classes with different degrees of susceptibility to the disease (see [20] for a general reference). In this case, the persistence or control of the disease depends on the initial size of the subpopulations. However, for reasonable estimates of the model parameters for influenza, and indeed for any infection where immunity acquired by natural infection does not wane significantly faster than that acquired by vaccination, we have shown that bistability does not occur, and therefore the disease can be controlled if and only if the basic reproductive number ($\mathscr{R}_0$) is reduced to values less than unity. This can be achieved only if infected individuals are not continuously introduced into the population and if the vaccination rate (ξ) and vaccine efficacy (σ) simultaneously exceed the thresholds ξ_c and σ_c, respectively (see (3.5) and (3.3)). However, if infected individuals are continuously recruited, no amount of vaccination would be enough to eradicate the disease. Increasing the level of vaccination nonetheless will always reduce the level of epidemicity of the disease, and vaccination can still be used to prevent a severe epidemic.

Due to the fact that typical outbreaks of influenza in a localized population are mostly caused by replication and spread of a single virus strain [1] (see also Table 1), the model presented here considers the dynamics of influenza transmission involving a single strain. If an outbreak occurs due to simultaneous person-to-person transmission of two or more viral strains (both of which are unaffected by partial cross-immunities), the model predicts a similar dynamics of influenza infection in the population. Further, in the event of two concurrent outbreaks, their dynamics would be similar, provided the viral strains involved are not subject to partial cross-immunity, which would affect the transmissibility of the strains [25, 32]. Although this fact is not considered in our study, the model presented here can be extended to monitor the effect of partial cross-immunity on the influenza dynamics involving two or more strains. It can also be extended to explore the long-term multiseason dynamics of influenza infection by employing time dependent parameters to deduce vaccination timing strategies (also known as pulse vaccination; see [23]).

Our model provides the vaccination rate, with a vaccine of known efficacy, necessary to control the spread of influenza in a population. This rate is determined based on the duration of infectiousness and the rate of contact between infected and susceptible individuals leading to the infection. This information is crucial for public health implementation of influenza control measures with the aid of a partially effective vaccine.

Appendix. Here, we shall show that recruitment of individuals into the vaccinated class $\tilde{V}$ does not affect the dynamics of the model, as long as $\xi > 0$. To do so, we consider the model with sources in all subpopulations as follows:

$$\frac{dS}{dt} = (1 - \epsilon_V - \epsilon_R - \epsilon_I) - \beta SI - \xi S - S + \omega V + \delta R, \tag{A.1}$$

$$\frac{dV}{dt} = \epsilon_V + \xi S - (1 - \sigma)\beta VI - (1 + \omega)V, \tag{A.2}$$

$$\frac{dI}{dt} = \epsilon_I + \beta SI + (1 - \sigma)\beta VI - (1 + \alpha)I, \tag{A.3}$$

$$\frac{dR}{dt} = \epsilon_R + \alpha I - (1 + \delta)R. \tag{A.4}$$

Making the assumption that the total population has reached its limiting value, so that $R = 1 - S - V - I$, gives the reduced model

$$\frac{dS}{dt} = (1 - \epsilon_V - \epsilon_I - \epsilon_R) - \beta SI - \xi S - S + \omega V + \delta(1 - S - V - I), \tag{A.5}$$

$$\frac{dV}{dt} = \epsilon_V + \xi S - (1 - \sigma)\beta VI - (1 + \omega)V, \tag{A.6}$$

$$\frac{dI}{dt} = \epsilon_I + \beta SI + (1 - \sigma)\beta VI - (1 + \alpha)I. \tag{A.7}$$

Using the change of variables

$$\begin{aligned} S &= \kappa S_1 - \frac{\epsilon_V}{\xi}, \\ V &= \kappa V_1, \\ I &= \kappa I_1, \\ R &= \kappa R_1, \\ t &= \mu t_1 \end{aligned}$$

gives

$$\begin{aligned} \frac{dS_1}{dt_1} &= (1 - \epsilon_I - \epsilon_R + \delta)\frac{\mu}{\kappa} + (1 + \delta)\frac{\epsilon_V \mu}{\kappa\xi} + (1 - \delta - \xi)\mu S_1 \\ &\quad + (\omega - \delta)\mu V_1 + \left(\frac{\beta\epsilon_V\mu}{\xi} - \delta\right) I_1, \end{aligned} \tag{A.8}$$

$$\frac{dV_1}{dt_1} = \mu\xi S_1 - (1 - \sigma)\mu\kappa\beta V_1 I_1 - \mu(1 - \omega)V_1, \tag{A.9}$$

$$\frac{dI_1}{dt_1} = \frac{\epsilon_I \mu}{\kappa} + \beta\mu\kappa S_1 I_1 + (1 - \sigma)\beta\mu\kappa V_1 I_1 - \frac{\beta\epsilon_V\mu}{\xi} I_1 - (1 + \alpha)\mu I_1. \tag{A.10}$$

Setting the sum of the constant coefficients in (A.8)–(A.10) to $\mu(1+\delta)$ allows one to solve for κ, giving

$$\kappa = 1 + \frac{\epsilon_V}{\xi} - \frac{\epsilon_R}{1 + \delta}. \tag{A.11}$$

Similarly, setting the coefficient of I_1 in (A.8) equal to $1-\mu(1+\delta)$ gives the following equation for μ:

$$\mu = \frac{\xi}{\xi + \beta\epsilon_V}. \tag{A.12}$$

Now redefining the constants

$$\begin{aligned}
1+\delta_1 &= \mu(1+\delta),\\
1+\omega_1 &= \mu(1+\omega),\\
\xi_1 &= \mu\xi,\\
\beta_1 &= \mu\beta\kappa,\\
1+\alpha_1 &= \mu\left(1+\alpha+\beta\frac{\epsilon_V}{\xi}\right),\\
\epsilon_I' &= \frac{\epsilon_I\mu}{\kappa}
\end{aligned}$$

yields

$$\frac{dS_1}{dt_1} = (1+\delta_1-\epsilon_I') - S_1(1+\delta_1+\xi_1) - \beta_1 S_1 I_1 + (\omega_1-\delta_1)V_1 - \delta_1 I_1, \tag{A.13}$$

$$\frac{dV_1}{dt_1} = \xi_1 S - (1-\sigma)\beta_1 V_1 I_1 - (1+\omega_1)V_1, \tag{A.14}$$

$$\frac{dI_1}{dt_1} = \epsilon_I' + \beta_1 S_1 I_1 + (1-\sigma)\beta_1 V_1 I_1 - (1+\alpha_1)I_1. \tag{A.15}$$

Therefore, these transformations change the system into the form of (2.10)–(2.12). Thus the model (A.5)–(A.7) with sources in all subpopulations can be reduced to a model with sources only in I and S, albeit with different parameters.

Acknowledgment. The authors would like to thank the referees for their constructive comments, which have improved the paper.

REFERENCES

[1] V. Andreasen, *Dynamics of annual influenza A epidemics with immuno-selection*, J. Math. Biol., 46 (2003), pp. 504–536.

[2] F. Ansaldi, P. D'Agaro, D. De Florentiis, S. Puzelli, Y. P. Lin, V. Gregory, M. Bennet, I. Donatelli, R. Gasparini, P. Crovari, A. Hay, and C. Campello, *Molecular characterization of influenza B viruses circulating in northern Italy during the* 2001-2002 *epidemic season*, J. Med. Virol., 70 (2003), pp. 463–469.

[3] R. M. Anderson and R. M. May, *Infectious Diseases of Humans*, Oxford University Press, London, New York, 1991.

[4] V. Andreasen, J. Lin, and S. A. Levin, *The dynamics of cocirculating influenza strains conferring partial cross-immunity*, J. Math. Biol., 35 (1997), pp. 825–842.

[5] E. A. Blumberg, C. Albano, T. Pruett, R. Isaacs, J. Fitzpatrick, J. Bergin, C. Crump, and F. G. Hayden, *Immunogenicity of influenza virus vaccine in solid organ transplant recipients*, Clin. Infect. Dis., 22 (1996), pp. 295–302.

[6] H. F. Berg, J. van Gendt, G. F. Rimmelzwaan, M. F. Peeters, and P. van Keulen, *Nosocomial influenza infection among post-influenza-vaccinate patients with severe pulmonary diseases*, J. Infec., 42 (2003), pp. 129–132.

[7] F. Brauer and P. van den Driessche, *Models for transmission of disease with immigration of infectives*, Math. Biosci., 171 (2001), pp. 143–154.

[8] S. Busenberg and P. van den Driessche, *Analysis of a disease transmission model in a population with varying size*, J. Math. Biol., 28 (1990), pp. 257–270.

[9] N. J. Cox and K. Subbarao, *Global epidemiology of influenza: Past and present*, Annu. Rev. Med., 51 (2000), pp. 407–421.

[10] N. J. Cox and K. Subbarao, *Influenza*, Lancet, 354 (1999), pp. 1277–1282.

[11] Y. Deguchi and Y. Tagasugi, *Efficacy of influenza vaccine in the elderly: Reduction in risk of mortality and morbidity during an influenza A ($H3N2$) epidemic for the elderly in nursing homes*, Int. J. Clin. Lab. Res., 30 (2000), pp. 1–4.

[12] L. Dorrell, I. Hassan, S. Marshall, P. Chakraverty, and E. Ong, *Clinical and serological responses to an inactivated influenza vaccine in adults with HIV infection, diabetes, obstructive airways disease, elderly adults and healthy volunteers*, Int. J. STD AIDS, 8 (1997), pp. 776–779.

[13] D. J. D. Earn, J. Dushoff, and S. A Levin, *Ecology and evolution of the flu*, Trends Ecol. Evol., 17 (2002), pp. 334–340.

[14] M. L. Garly and P. Aaby, *The challenge of improving the efficacy of measles vaccine*, Acta Tropica, 85 (2003), pp. 1–17.

[15] I. Gorotto, Y. Mandel, M. S. Green, N. Varsano, M. Gdalevich, I. Ashkenazi, and J. Shemer, *Influenza vaccine efficacy in young, healthy adults*, Clin. Infect. Dis., 26 (1998), pp. 913–917.

[16] B. T. Grenfell and R. M. Anderson, *Pertussis in England and Wales: An investigation of transmission dynamics and control by mass vaccination*, Proc. Roy. Soc. London Ser. B, 236 (1989), pp. 213–252.

[17] E. Hak, A. W. Hoes, and T. J. Verheij, *Influenza vaccination: Who needs them and when*, Drugs, 62 (2002), pp. 2413–2420.

[18] H. W. Hethcote, *The mathematics of infectious diseases*, SIAM Rev., 42 (2000), pp. 599–653.

[19] W. Janaszek, N. J. Gay, and W. Gut, *Measles vaccine efficacy during an epidemic in* 1998 *in the highly vaccinated population of Poland*, Vaccine, 21 (2003), pp. 473–478.

[20] C. M. Kribs-Zaleta, *Center manifolds and normal forms in epidemic models*, in Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction, IMA Vol. Math. Appl. 125, Springer-Verlag, New York, 2002, pp. 269–286.

[21] M. Koopmans, B. Wilbrink, M. Conyn, G. Natrop, H. Vander Nat, H. Venne, A. Meijer, J. van Steenbergen, R. Fouchier, A. Osterhaus, and A. Bosman, *Transmission of $H7N7$ avian influenza A virus to human beings during a large outbreak in commercial poultry farms in The Netherlands*, Lancet, 363 (2004), pp. 587–593.

[22] R. A. Lamb, *Genes and proteins of the influenza virus*, in The Influenza Viruses, R. M. Krug, H. Fraenkel-Conrat, and R. R. Wagner, eds., Plenum, New York, 1989, pp. 1–87.

[23] J. Lin, V. Andreasen, and S. A. Levin, *Dynamics of influenza A drift: The linear three-strain model*, Math. Biosci., 162 (1999), pp. 33–51.

[24] W. M. Liu and S. A. Levin, *Influenza and some related mathematical models*, in Applied Mathematical Ecology, Springer-Verlag, New York, 1989, pp. 235–252.

[25] S. A. Levin, J. Dushoff, and J. B. Plotkin, *Evolution and persistence of influenza A and other diseases*, Math. Biosci., 188 (2004), pp. 17–28.

[26] J. E. McElhaney, B. L. Beattie, R. Devine, R. Grynoch, E. L. Toth, and R. C. Bleackley, *Age-related decline in interleukin* 2 *production in response to influenza vaccine*, J. Amer. Geriatr. Soc., 38 (1990), pp. 652–658.

[27] S. M. Moghadas, *Modelling the effect of imperfect vaccines on disease epidemiology*, Discrete Contin. Dyn. Syst. Ser. B, 4 (2004), pp. 999–1012.

[28] A. S. Monto, K. Hornbuckle, and E. Ohmit, *Influenza vaccine effectiveness among elderly nursing home residents: A cohort study*, Amer. J. Epid., 154 (2001), pp. 155–160.

[29] B. R. Murphy and R. M. Chanock, *Immunization against virus disease*, in Fields Virology, 3rd ed., B. N. Fields, D. M. Knipe, P. M. Howley, et al., eds., Lippincott-Raven Publishers, Philadelphia, 1996, pp. 467–495.

[30] B. R. Murphy and R. G. Webster, *Orthomyxoviruses*, in Fields Virology, 3rd ed., B. N. Fields, D. M. Knipe, P. M. Howley, et al., eds., Lippincott-Raven Publishers, Philadelphia, 1996, pp. 1397–1445.

[31] C. W. Potter, *Chronicle of influenza pandemic*, in Textbook of Influenza, K. G. Nicholson, R. G. Webster, and A. J. Hay, eds., Blackwell Science, Oxford, UK, 1998, pp. 3–18.

[32] M. Nuño, Z. Feng, M. Martcheva, and C. Castillo-Chavez, *Dynamics of Two-Strain Influenza with Isolation and Partial Cross-Immunity*, Technical report BU-1606-M, BSCB Department, Cornell University, Ithaca, NY, 2002.

[33] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1996.

[34] S. F. Regan and C. Fowler, *Influenza. Past, present and future*, J. Gerontol. Nurs., 28 (2002), pp. 30–37.

[35] A. Scherer and A. R. McLean, *Mathematical models of vaccination*, Brit. Med. Bull., 62 (2002), pp. 187–199.

[36] H. R. Thieme and J. Yang, *An endemic model with variable re-infection rate and applications to influenza*, Math. Biosci., 180 (2002), pp. 207–235.

[37] WHO, http://www.who.int/csr/don/2004_01_21/en/print.html.

[38] Health Canada, http://www.hc-sc.gc.ca/pphb-dgspsp/fluwatch/03-04/w02_04/index.html.

[39] R. J. Webby and R. G. Webster, *Are we ready for pandemic influenza?*, Science, 302 (2003), pp. 1519–1522.

[40] J. M. Wood, *Developing vaccines against pandemic influenza*, Philos. Trans. Roy. Soc. London B Biol. Sci., 356 (2001), pp. 1953–1960.

[41] X. Xu, C. B. Smith, B. A. Mungall, S. E. Lindstrom, H. E. Hall, K. Subbarao, N. J. Cox, and A. Klimov, *Intercontinental circulation of human influenza $A(H1N2)$ reassortant viruses during the* 2001-2002 *influenza season*, J. Infec. Dis., 186 (2002), pp. 1490–1493.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 525–573

On Energy Surfaces and the Resonance Web[*]

Anna Litvak-Hinenzon[†] and Vered Rom-Kedar[‡]

Abstract. A framework for understanding the global structure of near-integrable n DOF Hamiltonian systems is proposed. To this aim two tools are developed—the energy-momentum bifurcation diagrams and the branched surfaces. Their use is demonstrated on a few near-integrable 3 DOF systems. For these systems possible sources of instabilities are identified in the diagrams, and the corresponding energy surfaces are presented in the frequency space and by the branched surfaces. The main results of this formulation are theorems which describe the connection between changes in the topology of the energy surfaces and the existence of resonant lower dimensional tori.

Key words. near-integrable Hamiltonians, parabolic resonances

AMS subject classifications. 70H08, 37J20

DOI. 10.1137/030600106

1. Introduction.

The study of the structure of energy surfaces of integrable systems and the study of resonances and instabilities in near-integrable systems developed into vast disparate research fields. The relation between the two received very little attention. Indeed, near regular level sets of the integrable Hamiltonian, the standard Arnold-resonance web structure appears, and the relation between the two fields reduces to the study of Arnold conjecture regarding instabilities in phase space. Here, we demonstrate that near singular level sets of the integrable Hamiltonian much information regarding possible instabilities of the near-integrable case may be deduced from the structure of the energy surface and its relations with resonance surfaces. We suggest that by adding some information to the traditional energy-momentum diagrams [7], which we name energy-momentum bifurcation diagrams (EMBD), one achieves a global qualitative understanding of the near-integrable dynamics. We relate the geometric properties of the surfaces corresponding to lower dimensional tori in this diagram to both bifurcations in the energy surface topology and the appearance of lower dimensional resonant tori.

Recall that energy surfaces of generic integrable Hamiltonian systems are foliated almost everywhere by n-tori,[1] which may be expressed locally as a product of n circles on which the dynamics reduces to simple rotations (the action-angle coordinates). A given compact regular level set (the set of phase space points with given values of the constants of motion) may be composed of several such tori. The energy surface is composed of all level sets with the same

[*] Received by the editors March 23, 2003; accepted for publication (in revised form) by T. Kaper May 24, 2004; published electronically November 9, 2004.
http://www.siam.org/journals/siads/3-4/60010.html

[†] Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK (litvak@maths.warwick.ac.uk). This author was partially supported by the Marie Curie Individual Fellowship HPMF-CT-2002-01742 and an INTAS grant.
[‡] Faculty of Mathematical and Computer Science, Weizmann Institute, Rehovot 76100, Israel (vered@wisdom.weizmann.ac.il). This author acknowledges the support of The MINERVA Foundation.

[1] For simplicity we consider here only the case of compact level sets.

energy. If these iso-energy level sets have different numbers of components, then there exists a singular level set on this energy surface which is not smoothly conjugate to a collection of n-tori. Under quite general conditions, Lerman and Umanskii [33] show that by using the reduction procedure [1] and Nehoroševresults [42], such a connected singular level set may be expressed locally as an $n-s$ dimensional torus with frequencies depending on $n-s$ actions crossed with a fixed point and its asymptotic manifolds in the remaining s DOF subsystem ($s \leq n$). This s DOF subsystem is called the normal system and its structure generally depends on the $n-s$ actions. (See exact statements in the formulation section below. For a complete treatment see [33].) The regular n-tori correspond to the case $s=0$. The larger the s the more cases and possibilities one has for the behavior in the normal directions. The larger the $n-s$, the more possibilities to transfer between the different cases, namely, to encounter bifurcations. In this context, the Lerman and Umanskii work [33] is mainly concerned with $n=s=2$, and the Lerman work [29] is mainly concerned with $n=s=3$ (yet it includes the generic bifurcation diagrams for the $s=1,2$ cases), whereas the works of Oshemkov [43], Fomenko [16], and Bolsinov and Fomenko [4] are mainly concerned with $n=2$, $s=1$. Here, we consider $s=1$ and $n \geq 2$, studying the implications of phase space bifurcations and resonances as a source for instabilities in the near-integrable systems.

Our main result, Theorem 2 (section 8.2), roughly states that *the existence of a nondegenerate $n-1$ dimensional torus of fixed points implies bifurcations in the topology of the energy surface*. Furthermore we prove that such a torus appears as an extrema of certain surfaces in the EMBD. In other words, this provides a relation among energy surface topology, bifurcations in the EMBD, and lower dimensional resonant tori.

The local coordinate representation of the lower dimensional singular tori naturally leads to investigation of critical points of the Hamiltonian function in the normal plane, with the remaining $n-s$ actions viewed as parameters [42]. Hence, as shown in [33, 29] and more recently in [25], for small $n-s$, singularity theory may be used to classify all generic bifurcations (or all generic bifurcations under given symmetries [25]) in the s DOF subsystem. Here we consider, in addition to the above bifurcations, extrema in the action variables of the Hamiltonian function evaluated along the singularity surfaces. We relate these extrema to strong resonances. A complete classification of all generic scenarios (of combinations of resonances and bifurcations in the normal plane), using singularity theory, is yet to be developed.

Since the integrable system has n integrals of motion, a representation of the energy surfaces corresponds to indicating the range of allowed motion and its character in some n dimensional space (the innocent words "its character" hide a vast body of work dedicated to understanding the topology of the level sets which are represented as points in this reduced space as discussed below). Traditional spaces for such representation are the frequency domain [27], the space of constants of motion (e.g., [2, 33, 40, 18, 17]), the energy-momentum space, and the momentum space (e.g., [2, 1, 6, 7, 47, 10, 33, 40]). Such presentations are all equivalent near regular level sets, where action-angle coordinates may be introduced. Furthermore, each of these representations is inherently nonunique as one may choose any nonsingular vector function of the conserved quantities to serve as the new set of coordinates. We propose that a convenient representation appears in a specific combination of energy and momentum space. Convenient here means the following.

C1. The geometric presentation supplies a concise summary of all the dynamics and geometrical features of the integrable system for all energy levels.

C2. The geometric presentation provides clear criteria for the location of special regions in phase space which are expected to produce strong instabilities under a given form of a perturbation.

While many presentations in the n dimensional space satisfy the first criteria, it appears that the second one has not been explicitly addressed. Notice that the first criteria deals with the integrable part of the Hamiltonian only. On the other hand, the second one depends on the nature of the applied perturbation; hence *the choice of the most convenient representation of the integrable system depends on the form of the perturbation*. These issues are explained more fully in section 4, where we propose a choice of convenient coordinates. A similar approach, in which the perturbation determines the appropriate integrable system, is taken in the partial averaging procedure. Other related works, in which the geometry of the energy surfaces and their intersection with the resonance web are related to the perturbed dynamics, are those on "resonance streaming," where it is argued that in 2 DOF integrable systems a small angle between the intersecting resonance and energy curves (plotted in the frequency space) enhances the effect of added noise [34, 48].

The various representations in the n dimensional spaces of the constants of motion identify the regions of the allowed motion but do not, in general, supply information regarding the topology of the level sets. Indeed, the classification of all the possible topologies of the level sets of energy surfaces of integrable Hamiltonians is extremely challenging (see [2, 33, 40, 17, 1, 47]) and has been completed for the 2 DOF case only [17, 33]. Lerman and Umanskii use the $n = 2$ integrals of motion near fixed points to obtain local and global information regarding the level sets of the integrable motion and to classify all possible generic homoclinic connections which are induced by the local behavior [30, 31, 32]. We use their formulation in our treatment of lower dimensional tori. Fomenko and coworkers suggested using graphs to represent all topologically distinct tori which appear in the integrable dynamics [17, 18, 16, 43, 4]. Some of these ideas have been extended to classify integrable 3 DOF dynamics such as the motion of rigid body [17, 10, 11]. Oshemkov, Fomenko, and coworkers use the 2 DOF constructions on given level sets of the third integral to analyze such systems [43, 17, 18], though Fomenko had also formulated a higher dimensional generalization of his theory [17]. Dullin et al. (see, e.g., [10, 11, 51] and references therein) have shown that such approaches may be used to develop schemes for computing action-angle coordinates even when the topology of the energy surfaces is complicated and finding n topologically independent circles (n circles which are irreducible to each other) is an a priori complicated task. Here we investigate the structure of energy surfaces with *very simple* topological structure for which we are able to generalize Fomenko–Oshemkov graphs to branched surfaces. These branched surfaces may be viewed as simple examples of Fomenko's higher dimensional theory.

The paper is ordered as follows: in section 2 we describe the type of the near-integrable Hamiltonians we study, with prototype examples of 3 DOF systems which demonstrate the appearance of nontrivial energy surfaces. In section 3 we formulate the notions of branched surfaces, topological bifurcations of the energy surfaces, and the EMBD. In section 4 we propose a convenient choice of momentum and explain how the choice of suitable coordinates depends on the form of the perturbation. In section 5 we describe the structure of the energy surfaces and the resonance web in the frequency space and in the energy momentum space near normally elliptic lower dimensional tori. The integrable structure in this a priori stable case is trivial and we add essentially no new insights to the known results. It is included here to build intuition for the next two sections. In section 6, we describe these structures near level sets corresponding to normally hyperbolic invariant $(n-1)$-tori. In section 7 we proceed to describe these structures near normally parabolic tori. In section 8 we prove our main theorems relating resonances and bifurcations. We conclude with a discussion section.

2. Formulation.

Consider a near-integrable Hamiltonian $H(q,p;\varepsilon) = H_0(q,p)+\varepsilon H_1(q,p;\varepsilon)$, $\varepsilon \ll 1$, $(q,p) \in M \subseteq \mathbb{R}^n \times \mathbb{R}^n$, where M is a $2n$ dimensional smooth symplectic manifold and H_0, H_1 are C^∞ (smooth).[2] We further assume that there exists $\varepsilon_0 > 0$ such that H_1 is bounded[3] for all real values of $(q,p;\varepsilon)$ for $\varepsilon \in [0,\varepsilon_0]$. H_0 represents the completely integrable part of the Hamiltonian (the unperturbed system) and its structure is described below. For any ε, a perturbed orbit with energy h resides on the energy surface $H_0(\cdot) = h - \varepsilon H_1(\cdot;\varepsilon)$. Hence, by assumption, the structure of the unperturbed energy surfaces and their resonance webs in an $O(\varepsilon)$-interval of energies near h supplies global information on the allowed range of motion of the perturbed orbits.

The integrable n DOF Hamiltonian, $H_0(q,p)$; $(q,p) \in M \subseteq \mathbb{R}^n \times \mathbb{R}^n$, has n integrals of motion, $H_0 = F_1, F_2, \dots, F_n \in C^\infty(M)$, which are functionally independent at almost all points of M and are pairwise in involution: $\{F_i, F_j\} = 0$; $i,j = 1,\dots n$. Assume that $n \geq 3$ and that the Hamiltonian level sets, $M_g = \{(q,p) \in M, \quad F_i = g_i; \quad i = 1,\dots,n\}$, are compact (this assumption implies, in particular, that the set $F_1, F_2, \dots, F_n$ is complete; see [33]). By the Liouville–Arnold theorem (see [42] and [2, 26]), the connected compact components of the level sets M_g, on which all of the dF_i are (pointwise) linearly independent, are diffeomorphic to n-tori, and hence a transformation to action-angle coordinates ($H_0 = H_0(I)$) near such level sets is nonsingular. Consider a neighborhood of a level set M_{g0} which possibly contains a singularity set at which s of the dF_i's are linearly dependent; on each connected and closed component of such a Hamiltonian level set there is some neighborhood D, in which the Hamiltonian $H_0(q,p)$ may be transformed by the reduction procedure to the form (see [33], [42])

$$H_0(x,y,I), \qquad (x,y,\theta,I) \in U \subseteq \mathbb{R}^s \times \mathbb{R}^s \times \mathbb{T}^{n-s} \times \mathbb{R}^{n-s}, \tag{2.1}$$

which does not depend on the angles of the tori, θ. The symplectic structure of the new integrable Hamiltonian (2.1) is $\sum_{j=1}^{s} dx_j \wedge dy_j + \sum_{i=1}^{n-s} d\theta_i \wedge dI_i$, where (x,θ,I) are the generalized action-angle variables ($s = 0$ corresponds to the maximal dimensional tori—the n-tori discussed above). The motion on the $(n-s)$ dimensional family (parameterized by the actions I) of $(n-s)$-tori is described by the equations

$$\dot{\theta}_i = \omega_i(x,y,I), \quad \dot{I}_i = 0.$$

The geometrical structure of the new Hamiltonian, $H_0(x,y,I)$, is such that for any fixed I ($I = (I_1,\dots,I_{n-s})$) an $(n-s)$-torus is attached to every point of the (x,y) plane (space, for

[2] One may relax these requirements to the C^r case, but this is left for future studies.

[3] It is probably sufficient to assume that for large (q,p) H_1 does not grow faster than H_0, namely, that for $\varepsilon_0 > 0$ sufficiently small there exist constants $C_1, C_2 \geq 0$ such that $|H_1(q,p;\varepsilon)| < C_1 + C_2\,|H_0(q,p)|$ for all $(q,p) \in M$ and $\varepsilon \in [0,\varepsilon_0]$, but we leave the details of this case for future work.

$s > 1$). The (x, y) plane (space) is called the *normal plane* (*space*) [2, 44, 3] of the $(n-s)$-tori and defines their stability type in the normal direction to the *family*[4] of tori. Invariant lower dimensional tori, of dimension $(n-l)$, generically exist for each $1 \leq l \leq n-1$; indeed, for any given s consider an m-resonant value of I. Then, for each such I, there exists an m dimensional family (corresponding to different initial angles) of $n-s-m$ dimensional tori. All these tori belong to the higher $n-s$ dimensional resonant torus, associated with I. The existence of such lower dimensional tori is restricted to the $n-s-m$ dimensional resonant surface of I values. A different type of lower dimensional invariant tori, which are of the main interest here, corresponds to isolated fixed point(s) of the s dimensional normal space. These appear on an $n-s$ dimensional manifold of I values, the singularity manifold (such a generalized fixed point corresponds to a manifold on which each dF_i for $i = 1, \ldots, s$ is linearly dependent on $dI_1, \ldots, dI_{n-s}$, where $dI_1, \ldots, dI_{n-s}$ are pointwise linearly independent). Locally, one may choose the (x, y, I) coordinate system so that for these tori

$$\nabla_{(x,y)} H_0(x, y, I)\big|_{p_f} = 0, \qquad p_f = (x_f, y_f, I_f). \tag{2.2}$$

Following the terminology of [33], the invariant tori on which (2.2) is satisfied are called here singular tori, and the manifolds of action values on which this equation is satisfied are called singularity manifolds. The structure of these is discussed in the next section.

Hereafter, consider the case $s = 1$ only. The invariant $(n-1)$-tori have an $(n-1)$ dimensional vector of inner frequencies, $\dot{\theta} = \omega(p_f)$. The normal stability type of such families of $(n-1)$-tori is determined by the characteristic eigenvalues (resp., Flouqet multipliers for the corresponding Poincaré map) of the linearization of the system about the tori; generically, these tori are either normally elliptic[5] or normally hyperbolic.[6] If the torus has one pair of zero characteristic eigenvalues in the direction of the normal (x, y) space, it is said to be *normally parabolic*. In addition, the *normal frequency*[7] [2, 44], Ω, of the $(n-1)$-tori is defined as the (nonnegative) imaginary part of the purely imaginary characteristic eigenvalues.[8]

Locally, in the (x, y, I) coordinate system, the normal stability of the invariant torus is determined by

$$\det\left(\frac{\partial^2 H_0}{\partial^2(x, y)}\bigg|_{p_f}\right) = -\lambda_{p_f}^2, \tag{2.3}$$

where p_f satisfies (2.2). Indeed, when λ_{p_f} is real and nonvanishing the corresponding family of tori is said to be normally hyperbolic, when it vanishes it is called normally parabolic, and when it is pure imaginary it is normally elliptic. For more details on the above see [33, 2, 5, 12, 13, 14, 15, 21, 22, 23, 24, 26, 44, 3] and references therein.

[4] Notice that a single torus belonging to this family has neutral stability in the actions directions. The normal stability referred to in the Hamiltonian context ignores these directions; see [5, 3] and references therein.

[5] If all the characteristic eigenvalues of an invariant lower dimensional torus (with respect to its normal (x, y) space) are purely imaginary (and do not vanish), it is said to be *normally elliptic*.

[6] If all the characteristic eigenvalues of an invariant lower dimensional torus (with respect to its normal (x, y) space) have a nonzero real part, it is said to be *normally hyperbolic*.

[7] In some references, these are called characteristic frequencies.

[8] In some references, e.g., [5], the normal frequencies are defined as the positive imaginary parts of the characteristic eigenvalues.

An example of a 3 DOF integrable Hamiltonian which is in the form (2.1), possesses families of invariant 2-tori of all three normal stability types (elliptic, hyperbolic, and parabolic) at $x = y = 0$, and satisfies all the stated above assumptions is

$$H_{bif}(x, y, I_1, I_2) = \frac{y^2}{2} - \frac{x^2}{2} I_1 + \frac{x^4}{4} + \left(\mu_1 + \frac{1}{2}\right) \frac{I_1^2}{2} + \frac{I_2^2}{2} + \alpha_2 I_2 + \alpha_3 I_1 I_2, \tag{2.4}$$

where α_2, α_3, and μ_1 are fixed parameters. Notice that this Hamiltonian has a Z_2 symmetry in the (x, y) coordinates. We will comment in the text when these symmetries play a role. In fact, the form of 3 DOF integrable Hamiltonian families in general position having a parabolic 2-tori has been derived (see [16],[33]). A complete study of their structure will be discussed elsewhere. We may compare it to a standard model of (Z_2 symmetric) a priori stable systems with bounded energy surfaces having a family of normally elliptic 2-tori at $x = y = 0$,

$$\begin{aligned} H_{st}(x, y, I_1, I_2) &= \frac{y^2}{2} + \frac{x^2}{2} + \frac{1}{8}\left(x^2 + y^2\right)^2 + \alpha_1 I_1 + \alpha_2 I_2 + \frac{I_1^2}{2} + \frac{I_2^2}{2} \\ &= \sum_{i=0}^{2} \left(\alpha_i I_i + \frac{I_i^2}{2}\right), \qquad \alpha_0 = 1, \end{aligned} \tag{2.5}$$

and to the corresponding (symmetric) a priori unstable system

$$H_{ust}(x, y, I_1, I_2) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4} + \alpha_1 I_1 + \alpha_2 I_2 + \frac{I_1^2}{2} + \frac{I_2^2}{2}, \tag{2.6}$$

which has a family of normally hyperbolic 2-tori at $x = y = 0$.

In sections 5, 6, 7, the representative models (2.5), (2.6), and (2.4) are studied in detail. In particular, for each of these models we construct the EMBD, find the branched surfaces which supply a representation for the energy surfaces geometry, and plot representing energy surfaces in the frequency plane. Below we define the branched surfaces and the EMBD. Some readers may find it helpful to read sections 5, 6, 7 first.

3. EMBD and branched surfaces. Fomenko and his coworkers have developed a sophisticated representation of integrable 2 DOF systems which leads to their orbital classification. Roughly, on each energy surface, they have suggested representing families of regular 2-tori by edges of a graph, whereas singular surfaces at which such families coalesce or undergo a change of orientation correspond to the vertices of the graph. The vertices corresponding to singular level sets are labeled according to the orbital changes they represent: “type A molecules” are vertices corresponding to normally elliptic circles, and “type B molecules” are vertices corresponding to a normally hyperbolic circle and its figure eight separatrices. Starred molecules correspond to change of orientation. To deal with higher dimensional systems, Fomenko and Oshemkov have suggested constructing tables of the 2 DOF molecules which list how these graphs vary as the constants of motion are changed. Other setups, closer to our construction, were suggested by Fomenko in [17] but, to the best of our knowledge, have not been examined or used for any specific model. Here, we explicitly construct the branched surfaces which we view as a different type of generalization of the simplest form (molecules A and B) of the Fomenko graphs.

Consider an integrable n DOF Hamiltonian on a $2n$ dimensional symplectic manifold M and its associated n integrals of motion $H_0 = F_1, F_2, \dots, F_n$. We call the set of constants of motion *valid* if they are *almost everywhere* functionally independent on M, are pairwise in involution, and are complete. Denote by A_h the set of allowed values of $F_2, \dots, F_n$ on the energy surface $E_h = \{(q,p) |\ H_0(q,p) = h\}$, namely, $A_h = \{(g_2, \dots g_n) |\ M_{(h,g_2,\dots g_n)} \neq \emptyset\}$ (recall that the level set M_g is defined as $M_g = \{(q,p) \in M, \quad F_i = g_i; \quad i = 1, \dots, n\}$). It follows that $E_h = \cup_{(g_2,\dots g_n)\in A_h} M_{(h,g_2,\dots g_n)}$. Let $k(g)$ denote the number of disconnected components of M_g, so $M_{(h,g_2,\dots g_n)} = \biguplus_{j=1}^{k(h,g_2,\dots g_n)} l_j^{(h,g_2,\dots g_n)}$, where l_j^g denotes a connected component of M_g. Fixing h, $k(h, g_2, \dots g_n)$ is constant when the level sets $M_{(h,g_2,\dots g_n)}$ deform smoothly with $g_2, \dots g_n$ (namely, at $h, g_2, \dots g_n$ the energy-momentum mapping is a trivial fiber bundle; see [2, 47]). Recall (see section 2) that the singularity surfaces of A_h are defined as the values of $(g_2, \dots g_n)$ for which there exists a point $(q,p) \in M_{(h,g_2,\dots g_n)}$ at which s of the dF_i are linearly dependent and the rank of the n vectors dF_i at the singularity is $n - s$. Let us denote the union of the singularity surfaces of some given A_h by A_h^S. For any finite range of (h, g) values, by the assumption on the compactness of the level sets, $k(h, g_2, \dots g_n)$ may change only across a singular level set, and hence

$$A_h^S \supseteq A_h^{GS} = \{(g_2, \dots g_n) |\ k(h, g_2, \dots g_n) \text{ is discontinuous in } g_2, \dots g_n\}.$$

Equality of these sets is expected in the generic case with $s = 1$. Nongeneric (e.g., symmetric) coincidences, by which disconnected level sets coincide and split at the same g value, may be similarly treated and will be ignored here (see [18] and [33] for discussion). The behavior of $k(h, g_2, \dots g_n)$ near singular level sets with $s \geq 2$ will be studied elsewhere.

We remark that Smale [47, 2] has called the values at which the energy-momentum map is not locally trivial in the differentiable sense (such as A_h^S) the bifurcation set—these values correspond to changes in the topology of the *level sets*. We follow here the Lerman and Umanskii terminology, referring to A_h^S as the singularity set, and we reserve the term bifurcation for changes in the energy surfaces structure (see Definition 3), which is the main focus of the current work.

Define a function $S_h : E_h \to \mathbb{R}^n$ as (recall that $g_1 = h$)

$$S_h(q,p) = (\delta(q,p), g_2, \dots g_n), \text{ where } F_i(q,p) = g_i, \ i = 1, \dots, n,$$

where the scalar function $\delta(q,p)$ satisfies the following:

- $\delta(q,p) = 0$ iff $k(g(q,p)) = 1$;
- two points belonging to the same level set have the same δ iff they belong to a connected component of M_g,

$$\{(q,p), (q',p') \in M_g \text{ and } \delta(q,p) = \delta(q',p')\} \Leftrightarrow (q,p), (q',p') \in l_j^g;$$

- $\delta(q,p)$ is smooth (C^r for some $r \geq 1$) for all $(q,p) \in E_h$ with $(g_2, \dots g_n) \in A_h \backslash A_h^S$, and $\delta(q,p)$ is continuous for $(g_2, \dots g_n) \in A_h$.

Hence, on each level set M_g, $\delta(q,p)$ (with $g = F(q,p)$) attains exactly $k(g)$ distinct values, i.e.,

$$\{\delta(q,p) |\ (q,p) \in M_g\} = \{\delta_1(g), \dots, \delta_{k(g)}(g)\}$$

and $\delta_i(g) \neq \delta_j(g)$ for $i \neq j$. Therefore, we may define

$$\delta(l_j^g) = \delta(q,p)|_{(q,p)\in l_j^g} = \delta_j(g),\ j = 1, \dots, k(g),$$

and hence

$$S_h(l_j^g) = S_h(q,p)|_{(q,p)\in l_j^g} = (\delta_j(g), g_2, \dots g_n).$$

Furthermore, if l_j^g and l_i^g coalesce to a (singular) level set $l_k^{g^*}$ as $g \to g^*$, then $\delta(l_{i,j}^g) \to \delta(l_k^{g^*})$. Summarizing, there is a 1-1 correspondence between the range of S_h and the connected components of the level sets in E_h, and this correspondence depends continuously on the phase space points even across singularities. In particular, if the level sets are compact, every point at which $S_h(q,p)$ is smooth corresponds to a single n-torus, and every point at which $S_h(q,p)$ is not smooth corresponds to a singular level set. Applying the above procedure for $n = 2$ leads immediately to the construction of the Fomenko graphs.

Definition 1. *The branched surface of an energy surface E_h is given by the surface $S_h(E_h)$; namely, it is the $n-1$ dimensional surface embedded in $\mathbb{R}^n$: $\mathcal{F} = \{y|\ y = S_h(q,p), (q,p) \in E_h\}$.*

Definition 2. *Two branched surfaces are equivalent if there exists a diffeomorphism of $\mathbb{R}^n$ which maps one branched surface to the other.*

Notice that, by definition, two equivalent branched surfaces are topologically conjugate. We require differentiable conjugacy so that the singular surfaces of two equivalent branched surfaces will be topologically conjugate to each other as well.

Figure 9 demonstrates that such a construction is possible for simple systems with singularities of order 1 (namely, with $s = 1$). These branched surfaces generalize the simplest molecules (A, B) of the Fomenko graphs. The application of this procedure to physical models is under current investigation; it may lead to further development of the theory, generalizing the other types of molecules which were constructed by Fomenko and his coworkers. Whether this will finally lead to a complete classification of integrable higher dimensional systems, as in the 2 DOF case, is unknown to us. First steps in this direction, in a more abstract setting, are included in [17]. In any case, it is easy to see the following.

Corollary 1. *Given an integrable Hamiltonian system, the branched surfaces constructed from two different valid sets of constants of motion (with $H_0 = F_1$) are equivalent.*

In particular, the construction of these surfaces from the EMBD, the frequency space diagram and the constant of motion diagrams are all equivalent, as demonstrated in sections 5, 6, 7.

Nearby energy surfaces correspond usually to a smooth deformations of each other; hence they will usually have equivalent branched surfaces. When nearby energy surfaces have different structures we say that the energy surface has undergone a bifurcation.

Definition 3. *h_c is an energy bifurcation point if the branched surfaces at h_c and at $h_c \pm \varepsilon$ are not equivalent for arbitrarily small ε.*

The EMBD supplies global information on the bifurcations of the energy surfaces structure and their relation to resonances; consider an integrable Hamiltonian system $H_0(q,p)$ in a region $D \subseteq M$ at which a transformation to the local coordinate system $H_0(x,y,I)$ with $s = 1$ is nonsingular. The energy-momentum map assigns to each point of the phase space (x,y,I) a point in the energy-momentum space $(h = H_0(x, y, I), I)$. The EMBD is a plot in the (h, I) space (for (h, I) in the range of D) which includes

- the region(s) of allowed motion (the closure of all regions in which the energy-momentum mapping is a trivial fiber bundle; see [2, 47]);
- the $s = 1$ singular surfaces $(h, I) = (H_0(p_f), I_f)$ (see (2.2)) where the normal stability of the corresponding singular $(n-1)$-tori, defined by (2.3), is indicated;
- the strongest resonance surfaces on which the inner frequencies of the tori vanish, $\omega_i = 0$, and the regions in which back-flow occurs (where $\dot{\theta}_i(x, y, I)$ changes sign along the level set $(h = H_0(x, y, I), I)$ for some $i = 1, \ldots, n-s$);
- the energies at which topological bifurcations occur and the branched surfaces in the intervals separate by these bifurcation points.

4. The dependence of the EMBD on the form of the perturbation.

The EMBD depends on the choice of the generalized action-angle co-ordinates (x, y, I). In particular, a symplectic transformation of the (θ, I) coordinates to other generalized action-angle coordinates (φ, J) may change some geometric properties of this diagram. Furthermore, transformations of the form $(\theta, I) \to (\varphi = \theta - \varpi(I)t, I)$, which correspond to moving relative to particular tori with frequency vectors $\omega = \varpi(I)$, for which $H_{new}(J) = H_0(J) - \overline{H}(J)$ (where $\varpi(I) = \frac{\partial \overline{H}(I)}{\partial I}$), may change the topology of the energy surfaces, the structure of their singularity manifolds, and the nature of their intersections with the resonance hypersurfaces. Hence, it appears that finding the "correct" representation is ill defined in the context of the integrable system. *We propose that the form of the perturbation resolves these issues.*

The role of the form of the perturbation is apparent when one considers the procedure of partial averaging in near-integrable Hamiltonian systems (see [2, p. 173]). First, this procedure implies that if for a given perturbation the number of independent resonant vectors, r, is smaller than the number of angle variables (here $n-1$), then, by averaging over the $n-r-1$ nonresonant directions one can obtain a system which is exponentially close to the $r+1$ dimensional system depending on the $n-r-1$ parameters (the averaged actions). Hence, with no loss of generality, we assume here that the perturbation includes $n-1$ independent resonant directions. Then, it follows that for autonomous systems the transformation $(\theta, I) \to (\varphi = \theta - \varpi(I)t, I)$ produces time-dependent Hamiltonians, and hence $\varpi(I)$ must vanish identically.[9] This trivial statement implies, in particular, that given an a priori stable near-integrable system of the form (2.5), for generic vector α, it may not be transformed to a system with $\alpha = 0$ without introducing time-dependent perturbations.

The second issue which arises is the choice of the action variables, which implicitly determines which resonances are considered strong. This issue is again well understood in the context of partial averaging [2]; the form of the anticipated *perturbation* determines which of the resonant surfaces will produce the strongest response; consider the near-integrable system expressed near a singular level set in some local generalized action-angle coordinates:

$$H(q, p) = H_0(x, y, I) + \varepsilon H_1(x, y, \theta, I). \tag{4.1}$$

[9] Here we need the assumption on having $n-1$ independent resonant relations. Indeed, if $n \geq 3$ and $H_1(x, y, I, \theta) = \cos(\theta_1 - \theta_2)$, one can immediately see that this statement is false.

Consider the Fourier series of H_1:

$$H_1(q,p) = H_1(x,y,\theta,I) = \sum_{|k|=|k_1|+\cdots+|k_n|\geq r_0} h_k(x,y,I)\exp(\mathrm{i}\langle k,\theta\rangle).$$

Then, provided h_k are monotonically decreasing (if H_1 is assumed to be analytic, h_k decays exponentially for large $|k|$, and if H_1 is assumed to be C^r, the decay rate is of order $|k|^{-r}$), the strongest resonances are given by the frequencies satisfying $\langle k,\omega\rangle = 0$ for k values which are included in the sum ($h_k \neq 0$ near the singular level set) and satisfy $|k| \approx r_0$; see [2] for an exact definition and discussion. In particular, by a change of the action angle coordinates $(\theta,I) \to (\varphi,J)$ one may arrange the terms so that the first n terms in the above sum have $k^j = r_j e_n^j = (0,\ldots,r_j,\ldots 0)$ for $j = 1,\ldots,n$ (where r_j is in the jth place, and are monotonically increasing with j). Then, the strongest resonances occur along the action variables directions.

Summarizing the above observations, we propose the following.

Definition 4. *The local generalized action-angle coordinates (x,y,θ,I) of the integrable part of a near-integrable system are called suitable if the following hold:*

- *The perturbed system is autonomous.*
- *The strongest $n-s$ resonant terms are nonzero and are aligned, in decreasing order, along the $n-s$ actions. More precisely, let*

$$H_1(q,p) = H_1(x,y,\theta,I) = \sum_j h_{k^j}(x,y,I)\exp(\mathrm{i}\langle k^j,\theta\rangle), \quad j \in \mathbb{N}, \tag{4.2}$$

with $|k^j|$ monotonically increasing with j and $\|h_{k^j}\|$ monotonically decreasing with j for equal $|k^j|$ values. Then, $k^j || e_{n-s}^j$ (i.e., k^j is parallel to e_{n-s}^j) for $j = 1,\ldots,n-s$, where e_{n-s}^j is the $n-s$ dimensional unit vector with 1 at the jth entry.

Notice that the sum in (4.2) is assumed to have *at least* $n-s$ terms with $n-s$ independent k^j vectors. In general the sum has an *infinite number* of nonzero terms, where the k^j vectors with $j > n-s$ are linearly dependent on $(k^1,\ldots,k^{n-s})$.

Lemma 1. *If (x,y,θ,I) and (q,p,φ,J) are two suitable coordinate systems for the Hamiltonian (4.1), then $I = J$ and $\varphi = \theta + f(I)$.*

Proof. First, we recall that by generalized action-angle coordinates we assume that near the singular level set the unperturbed equations of motion for (I,θ) are of the form

$$\frac{dI}{dt} = 0 = -\frac{\partial H_0}{\partial\theta}, \ \frac{d\theta}{dt} = \omega(x,y,I) = \frac{\partial H_0}{\partial I},$$

and similarly for the (J,φ) coordinates. Let $W(J,\theta)$ denote the generating function which transforms (x_f,y_f,θ,I_f) to (q_f,p_f,φ,J_f). Since both θ and φ are angle coordinates, it follows that $W(J,\theta)$ is of the form $W(J,\theta) = J^T R\theta + F(J)$, where R is a unimodular integral matrix (see [2, p. 173]), so that

$$\begin{aligned} I_f^T &= \frac{\partial W}{\partial\theta} = J_f^T R, \\ \varphi &= \frac{\partial W}{\partial J_f} = R\theta + F'(J_f). \end{aligned}$$

In particular, the requirement that both θ and φ are 2π periodic in all their arguments and that the transformation is smooth implies that R is a constant integral matrix, independent of J_f. Furthermore, extending the transformation to a neighborhood of the singular level set must still preserve this property. To complete the proof, we need to show that $R = Id$. Indeed, expressing the perturbed part of the Hamiltonian in the (x, y, θ, I) coordinates and using the above transformation in (4.2), we obtain

$$H_1(q, p, \varphi, J) = \sum_j \widetilde{h_{k^j}}(q, p, J) \exp(\mathrm{i} \left\langle k^j, R^{-1}\varphi \right\rangle), \quad j \in \mathbb{N},$$

with $k^j || e^j_{n-s}$ for $j = 1, \ldots, n-s$ (where R^{-1} is a unimodular integral matrix as well). Therefore, insisting that q, p, φ, J are suitable implies that for all φ

$$\left\langle e^j_{n-s}, R^{-1}\varphi \right\rangle = \left\langle e^j_{n-s}, \varphi \right\rangle \quad \text{for} \quad j = 1, \ldots, n-s$$

and hence that $R^{-1} = R = Id$. ■

Conversely, one may take the usual convention by which, given an integrable Hamiltonian in a given coordinate system, the analysis determines which form of the perturbation will cause the largest instability in the vicinity of a given resonance junction. In particular, by the appropriate change of coordinates of the integrable system, one obtains that the *strongest* resonances possible are realized when $k = e^j_n = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$. With this view the notion of strongest resonances is inherently coordinate-dependent.

In our presentation of the energy surfaces in the energy-momentum plots we relate changes in the energy surface singular structures to strong resonances of lower dimensional tori and to instabilities in the near-integrable system. To make such statements well defined, we insist that the coordinates we use are locally suitable coordinates. On the other hand, results which relate strong resonances to topological changes in the energy surfaces (e.g., Theorem 2) are independent of the choice of the coordinate system.

5. A priori stable systems. To obtain a good understanding of the proposed presentation of the EMBD we begin with the simplest and most familiar model of a priori stable systems near the lower dimensional torus. Let us examine the presentation of the regular part of the energy surface first in the frequency space and then in the energy-momentum space.

5.1. Energy surfaces in the frequency space (S). For the standard Hamiltonian H_{st} (see (2.5)) the transformation from momentum to frequency variables is a shift ($\omega(I) = \alpha + I, \quad \alpha = (1, \alpha_1, \alpha_2)$) and is regular everywhere, so we can write

$$H_{st}(I) = H_{st}(\omega) = \frac{1}{2}\left\|\omega(I)\right\|^2 - \frac{1}{2}\left\|\alpha\right\|^2, \tag{5.1}$$

and we obtain the standard result that in the definite case the energy surfaces appear in the frequency space as spheres centered at $\omega = 0$. The natural oscillations near the elliptic fixed point $x = y = 0$ (where the transformation from the (x, y) coordinates to the action-angle coordinates is singular) correspond to the circle $\omega_0 = \alpha_0 = 1$, and in this representation appear as a regular level set of the energy surface. Here it is natural to insist on positive I_0 value,

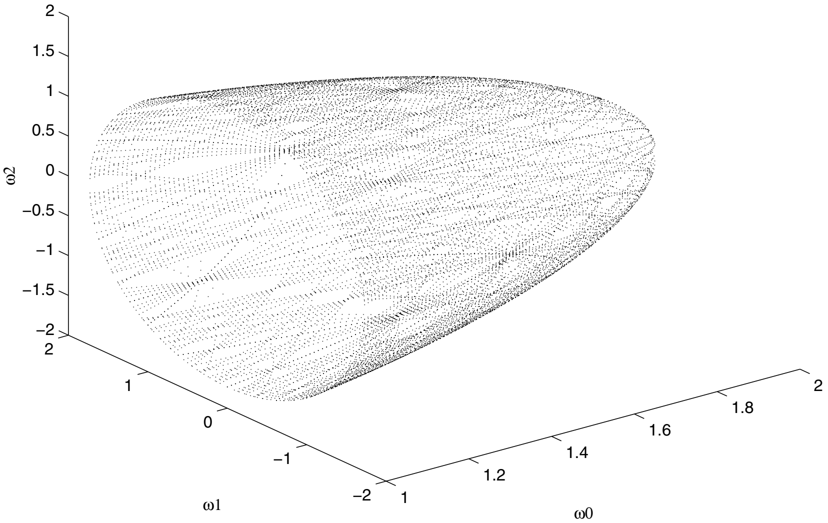


Figure 1. *A resonance web on a cap of an energy surface of an a priori stable system ((5.1) with $h = 0.5$).*

leading to energy surfaces in the form of "caps" with boundaries: $\omega^h = \{\omega | \ \|\omega\|^2 = 2h + \|\alpha\|^2, \omega_0 \geq \alpha_0 = 1\}$. The boundary $\omega_0 = \alpha_0$ corresponds to the family of lower dimensional tori $x = y = 0$ on the given energy surface; see Figure 1. In the figure we also show the dense intersection of the resonance surfaces, given by planes passing through the origin, with this cap (see [2, 23, 34] and references therein). The planes $\omega_i \equiv 0,\ i = 0, 1, 2$, correspond to the strongest resonances. Notice that the only energy surface which includes the origin is a sphere with diminishing radius, and such an energy surface is disallowed for systems of the form (5.1) since[10] $\|\alpha\| \neq 0$.

5.2. Energy surfaces in the energy-momentum space (S). In Figure 2, we construct the EMBD of the system (2.5) by presenting the energy surfaces in the space (H_0, I_1, I_2), where $H_{st}(x, y, I) = H_0$. For any given energy $H_0 = h$, the allowed region of motion is bounded by the family of normally elliptic 2-tori $(x, y, I) = (0, 0, I(h))$. The corresponding singularity surface in the EMBD is given by the paraboloid

$$p^0_{ell}(h, I_1, I_2) = \left\{ (h, I_1, I_2) | \quad H_{st}(0, 0, I) = \frac{1}{2} \sum_{i=1}^{2} \left((\alpha_i + I_i)^2 - \alpha_i^2 \right) = h,\ h \geq h_{ell} = 0 \right\};$$

namely, for a given h, (I_1, I_2) belong to a circle of radius $\sqrt{2h + \alpha_1^2 + \alpha_2^2}$ which is centered at $(-\alpha_1, -\alpha_2)$.The singularity manifolds corresponding to *normally elliptic* invariant tori are denoted by a collection of *solid* curves in the EMBD as demonstrated in Figure 2. To see that

[10] Recall that changing α corresponds to considering perturbations which are quasi-periodic functions of time.

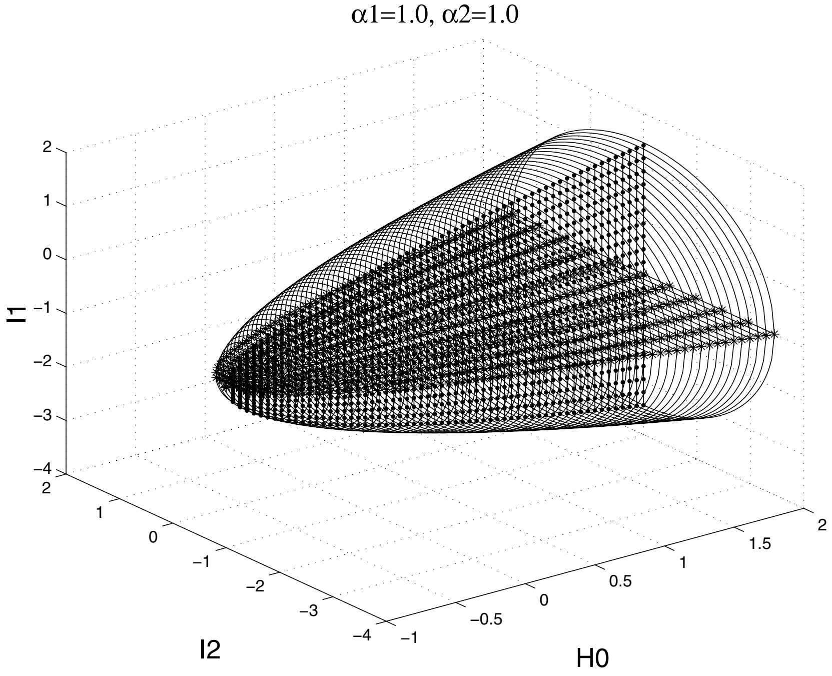


Figure 2. *EMBD of an a priori stable system.*

motion is allowed only for I values which are interior to this paraboloid notice that from (2.5)

$$0 \leq \frac{y^2}{2} + \frac{x^2}{2} + \frac{1}{8}\left(x^2+y^2\right)^2 = h - \frac{1}{2}\sum_{i=1}^{2}(I_i+\alpha_i)^2 + \frac{\alpha_1^2+\alpha_2^2}{2}.$$

The energy surfaces which appear as caps in the frequency space are "flattened" here to discs in the energy-momentum space. An example of an energy surface in the (I_2, I_1) plane, corresponding to the two dimensional (2D) slice of Figure 2 at $H_0 = 1$, is presented in Figure 3A. The thin vertical lines in the 2D sections of the EMBD indicate the region of allowed motion. In Figure 3B we present a 2D slice in the (H_0, I_1) plane at $I_2 = 0$, on which we schematically indicate the corresponding Fomenko graph by a thick black line. The Fomenko graph for any positive h and an interior I_2 value is simply a segment: each interior point on this segment corresponds to a single 3-torus, and each of the end points corresponds to a normally elliptic 2-torus ("atom A" in [18]). For such a fixed I_2 value the energy surface appears as a 2-sphere in the (x, y, I_1) space. The poles of this sphere are normally elliptic 2D tori, and they correspond to the boundaries of this component of the energy surface. Equivalently, we may think of a natural generalization of the Fomenko graphs to branched surfaces, and in this trivial case the branched surface is simply a single disk, as shown in Figure 9A: an interior point to the disc corresponds to a 3-torus, and a point on the disc boundary corresponds to a 2-torus.

The *strong resonances* $\omega_i = 0, i = 1, 2,$ correspond here to the hyperplanes $I_i = -\alpha_i$.

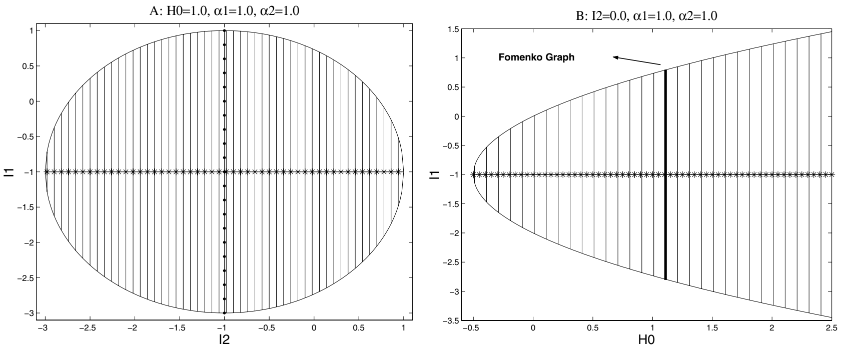


Figure 3. *2D slices of an EMBD of an a priori stable system.* A: *One energy surface.* B: *Range of energy values for a fixed value of* I_2 *and a schematic Fomenko graph.*

Their intersections with the singularity manifold p^0_{ell} satisfy

$$\dot{\theta}_i = \frac{\partial H}{\partial I_i}\bigg|_{x=y=0, I_i=-\alpha_i} = \frac{dH(p^0_{ell})}{dI_i}\bigg|_{x=y=0, I_i=-\alpha_i} = 0.$$

Namely, it *corresponds to a fold in the singularity manifold* p^0_{ell}. Here the relation between total derivatives with respect to I_i along p^0_{ell} and the corresponding partial derivatives of H is trivial. Notice that the same relation is satisfied even when the location of the singularity manifold depends on I_i since $\nabla_{x,y}H$ vanishes on the singularity surfaces. This latter property is coordinate-independent as long as the coordinates are suitable.

In Figures 2 and 3 we indicate the strongest resonant 3-tori by starred lines (for $\dot{\theta}_1 = 0$) and dotted lines (for $\dot{\theta}_2 = 0$). The intersection of these surfaces (here planes) with the singularity surface ($x = y = 0$) corresponds to the strongly resonant families of lower dimensional tori $\{(x, y, I) = (0, 0, -\alpha_1, I_2)\}$,$\{(x, y, I) = (0, 0, I_1, -\alpha_2)\}$. These two families of 2-tori intersect at the minimal possible energy where $(I_1, I_2) = (-\alpha_1, -\alpha_2)$, corresponding to a 2-torus of fixed points. Namely, this torus of fixed points corresponds to a topological change in the energy surface—for energies below the value at which this torus appears there is no allowed motion. For energies above it we have one connected component of energy surface as described above. This observation is a trivial manifestation of Theorem 2.

5.3. Qualitative behavior of the near-integrable system (S). The motion in the near-integrable system $H(\theta, I) = H_{st}(I) + \varepsilon H_1(\theta, I)$ is restricted to the energy level $H = h$. Since $H_1(\theta, I)$ is assumed to be bounded, we obtain that the unperturbed energy surfaces with $H_{st}(I) = h^*$, $|h^* - h| < C\varepsilon$ supply an a priori bound to the motion. For energy surfaces of large extent, such an a priori bound is irrelevant due to Nehorošev-type theorems and the Arnold conjecture. Hence, in this case the only new information obtained from the EMBD is regarding the appearance of strong lower dimensional resonant tori (and even this information can be extracted from the frequency plot). Near the doubly resonant normally elliptic lower dimensional torus, where $(x, y, I) = (0, 0, -\alpha_1, -\alpha_2)$, one may expect small perturbations to produce large instabilities. Our trivial observation regarding the extent of the energy surface immediately shows that the extent of the instability cannot be larger than $O(\sqrt{\varepsilon})$, the extent in the I space of the energy surfaces with $H_{st}(I) = h^*$, $|h^* - h| < C\varepsilon$. We may expect that the behavior near the doubly resonant torus will be dramatically different if the dependence on the actions is either linear or indefinite, e.g.,

$$H_{st-unbounded}(x, y, I_1, I_2) = \frac{y^2}{2} + \frac{x^2}{2} + \frac{1}{8}\left(x^2 + y^2\right)^2 + \alpha_1 I_1 + \alpha_2 I_2 - \frac{I_1^2}{2} + \frac{I_2^2}{2};$$

namely, the energy surfaces are unbounded, and, in particular, the energy surface passing through the elliptic double resonant fixed point is unbounded. Such considerations are the trivial analogues to the nonlinear stability theorems of Arnold–Marsden and have been studied and discussed in the context of normal forms near elliptic fixed points [2, 40].

6. A priori unstable systems.

The phase space structure of the standard Hamiltonian H_{ust} (see (2.6)),

$$H_{ust}(x, y, I_1, I_2) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4} + \frac{1}{2}\sum_{i=1}^{2}\left((\alpha_i + I_i)^2 - \alpha_i^2\right) \tag{6.1}$$

$$= H_{xy}(x, y) + H_I(I), \tag{6.2}$$

is given by the product of a figure eight motion in the xy plane and a family of 2-tori in the (θ, I) space. The precise structure of each energy surface, which demonstrates how a given energy may divide between the three modes (degrees of freedom) requires a bit more attention. Since there are no global action-angle coordinates in the (x, y) plane, it is instructive to start with the presentation of the energy surfaces in the energy-momentum space and then discuss the presentation in the frequency space.

6.1. Energy surfaces in the energy-momentum space (U).

To construct the EMBD we find the singularity manifolds of the Hamiltonian (6.1). These manifolds correspond to fixed points of $H_{xy}(x, y)$. The normally elliptic singularity surfaces, corresponding to $\{x = \pm 1, y = 0\}$, are given by the identical[11] paraboloids

$$p_{ell}^{\pm}(h, I_1, I_2) = \left\{(h, I_1, I_2)| \quad \frac{1}{2}\sum_{i=1}^{2}\left((\alpha_i + I_i)^2 - \alpha_i^2\right) = h + \frac{1}{4},\ h \geq h_{ell}\right\},$$

where

$$h_{ell} = -\frac{1}{4} - \frac{1}{2}\sum_{i=1}^{2}\alpha_i^2.$$

The normally hyperbolic singularity surface corresponding to $x = y = 0$ and its separatrices is given by the paraboloid

$$p_{hyp}^{0}(h, I_1, I_2) = \left\{(h, I_1, I_2)| \quad \frac{1}{2}\sum_{i=1}^{2}\left((\alpha_i + I_i)^2 - \alpha_i^2\right) = h,\ h \geq h_{hyp}\right\},$$

[11] Adding an asymmetric term like ηx to H_{ust} lifts this degeneracy.

where

$$h_{hyp} = -\frac{1}{2}\sum_{i=1}^{2}\alpha_i^2.$$

In Figure 4 an EMBD of system (6.1) is presented as two nested paraboloids. The singularity manifolds are drawn according to the normal stability of the lower dimensional invariant tori they represent—the normally elliptic singularity manifolds ($p_{ell}^{\pm}(h, I_1, I_2)$) are drawn as a collection of solid curves, whereas the *normally hyperbolic* singularity manifold ($p_{hyp}^{0}(h, I_1, I_2)$) is drawn as a collection of black *dashed* curves. Thus we follow the traditional notation in bifurcation diagrams.

Given an energy surface $H_{ust}(x, y, I_1, I_2) = H_0 = h$ with $h_{ell} \le h < h_{hyp}$, the three dimensional (3D) and 2D EMBD look locally similar to those of the a priori stable system, presented in Figures 2 and 3: for each fixed energy value in this range the energy surface is a disk in the (I_2, I_1) plane. However, each point interior to this disk corresponds to *two* sets of 3-tori, one in each well of the potential of $H_{xy}(x, y)$. Points on the boundary of the disk correspond to the *two* normally elliptic 2-tori, $\{x = \pm 1, y = 0, H_{ust}(\pm 1, 0, I_1, I_2) = h\}$. Hence, the Fomenko graph for any one dimensional (1D) section of each such disk is given by two disconnected segments, as shown in Figure 5B. Equivalently, the generalized branched surfaces for this range of energies is the union of two disconnected discs (see Figure 9B). Each point belonging to the interior of the branched surfaces represents, as before, a single 3-torus, and every point on the solid boundary of the discs represents, as before, a single, normally elliptic 2-torus. The multiplicity in the number of components of the level set corresponding to a given (h, I_1, I_2) is expressed by the multiplicity in the number of components of the branched surfaces for these values of (h, I_1, I_2); see section 3.

For $h \ge h_{hyp}$ the energy surfaces include the singular level set of the separatrices, which divides the energy surface into two topologically different regimes; see Figure 5A. A point (h, I_1, I_2) inside the disk enclosed by $p_{hyp}^{0}(h, I_1, I_2)$ (the dashed circle in Figure 5A) corresponds to a single 3-torus. Trajectories belonging to this torus encircle both wells in the xy plane. A point inside the ring bounded between $p_{hyp}^{0}(h, I_1, I_2)$ and $p_{ell}^{\pm}(h, I_1, I_2)$ corresponds to *two* sets of 3-tori; trajectories belonging to one of these tori oscillate in one of the wells in the xy plane. The Fomenko graph for this case is shown schematically on the cross-section in Figure 5B (in thick black). The generalized branched surfaces here are two rings which are glued together in a central disk (Figure 9C). Each regular point of the branched surface corresponds to a single 3-torus, each point belonging to the dashed circle corresponds to a normally hyperbolic 2-torus and its separatrices, and each point belonging to the solid (outer) boundaries of the rings corresponds to a single normally elliptic 2-torus.

Intersections of the singularity manifolds with the hypersurfaces of strongest resonances correspond to folds of these singularity manifolds in the EMBD (see Figures 4 and 5). This is the essence of Theorem 1 (see section 8.1). For example, the paraboloids $p_{hyp}^{0}(h, I_1, I_2)$ and $p_{ell}^{\pm}(h, I_1, I_2)$ fold as they cross the surface $I_1 = -\alpha_1$ (and similarly at $I_2 = -\alpha_2$) and indeed $\dot{\theta}_i|_{p_f} = \alpha_i + I_i$. Thus, the families of 2-tori, $p_{ell}^{\pm}(h, -\alpha_1, I_2)$, $p_{ell}^{\pm}(h, I_1, -\alpha_2)$, $p_{hyp}^{0}(h, -\alpha_1, I_2)$, and $p_{hyp}^{0}(h, I_1, -\alpha_2)$, are all resonant and the 2-tori $p_{ell}^{\pm}(h_{ell}, -\alpha_1, -\alpha_2)$ and $p_{hyp}^{0}(h_{hyp}, -\alpha_1,$

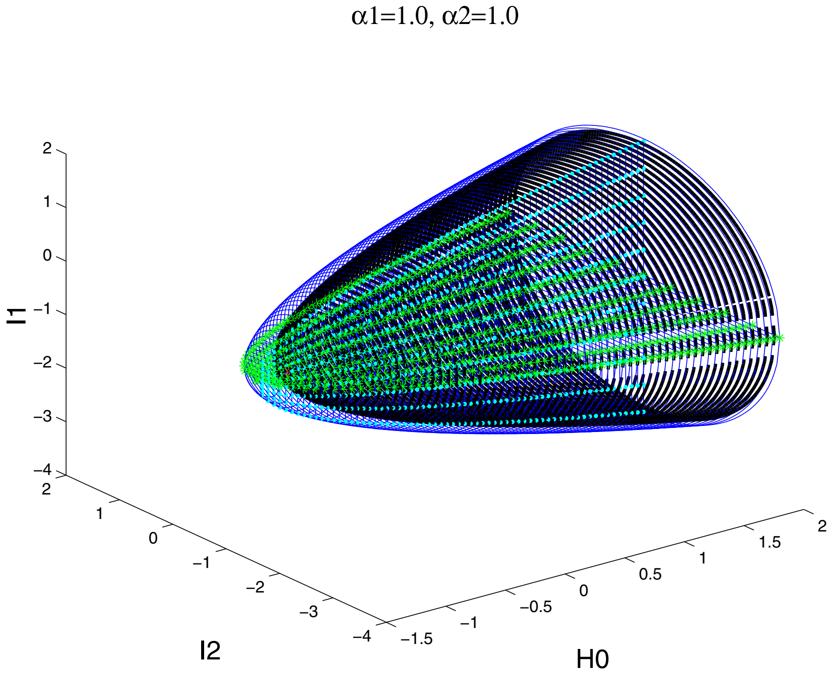


Figure 4. *EMBD of an a priori unstable system.*

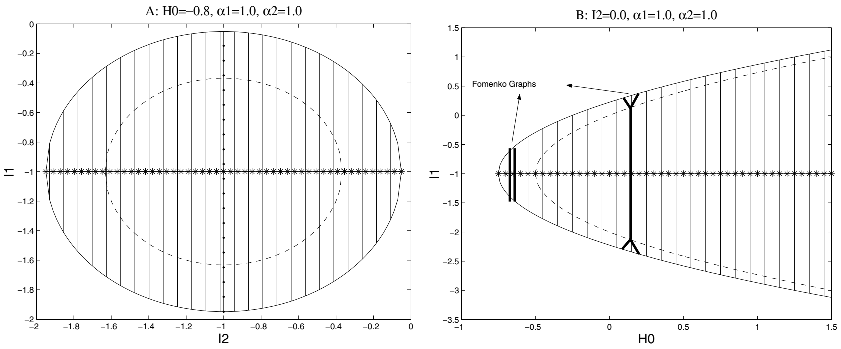


Figure 5. *2D slices of an EMBD of an a priori unstable system with $h > h_{hyp}$.* A: *One energy surface.* B: *Range of energy values for a fixed value of I_2 and a schematic Fomenko graph.*

$-\alpha_2$) are doubly resonant; these are 2-tori of fixed points. In Figure 4 the strong resonance in the I_1 direction ($\dot{\theta_1} = 0$) is denoted by a surface of green starred lines and the strong resonance in the I_2 direction ($\dot{\theta_2} = 0$) by a surface of cyan dotted lines; the *double fold* corresponding to a 2-resonant hyperbolic 2-torus (a hyperbolic torus of fixed points) is denoted by a red star.

Observe that the topology of the family of equi-energy normally hyperbolic lower dimensional tori ($p^0_{hyp}(h,\cdot)$, with fixed h) changes exactly at this double fold point, $p^0_{hyp}(h_{hyp}, -\alpha_1, -\alpha_2)$, where a 2-torus of fixed points resides; for $h < h_{hyp}$ the singularity surface $p^0_{hyp}(h,\cdot)$ does not exist and the energy surfaces have two disconnected components (Figure 9B), whereas for

$h > h_{hyp}$ the singularity surface $p^0_{hyp}(h,\cdot)$ is a circle and the two components of the energy surface connect on this circle (Figure 9C). This is again a manifestation of Theorem 2 (see section 8.1). Similarly, for the natural n DOF generalization of H_{ust},

$$H^n_{ust}(x,y,I_1,\dots,I_{n-1}) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4} + \frac{1}{2}\sum_{i=1}^{n-1}\left((\alpha_i+I_i)^2-\alpha_i^2\right) \tag{6.3}$$

$$= H_{xy}(x,y) + H^n_I(I), \tag{6.4}$$

$p^0_{hyp}(h,\cdot)$ changes from nonexistence for $h < h_{hyp}$ to an $n-2$ sphere for $h > h_{hyp}$. In particular, for $n = 2$, there are either none or two nonresonant hyperbolic circles on each energy surface. If the dependence on the I variables is indefinite, then one may have other topological changes in $p^0_{hyp}(h,\cdot)$ occurring at the $(n-1)$-resonant $(n-1)$-torus; either the genus of $p^0_{hyp}(h,\cdot)$ changes or two components of $p^0_{hyp}(h,\cdot)$ coalesce/separate. Notice that the flow along the lower dimensional $(n-1)$-torus reverses its direction as a strong resonance surface is crossed (namely, $\dot{\theta}_i|_{p_f}$changes its sign there). This property holds for the general case of nonseparable systems as well (see [46, 38, 37]).

We emphasize that the appearance of an $n-1$ dimensional torus of fixed points which is normally hyperbolic is a persistent[12] phenomena in integrable n $(n \geq 2)$ DOF Hamiltonian systems [38] and is not related to the symmetric form of (6.3).

6.2. Energy surfaces in the frequency space (U). For small energy levels, $h_{ell} \leq h < h_{hyp}$, we have seen that the disk $H_{ust}(x,y,I_1,I_2) = h$ in the (I_2,I_1) plane (see Figures 5B and 9B) corresponds to two separate smooth compact components of the energy surface. In the frequency space these appear as one cap of hyperbola, centered at the origin. Indeed, the natural frequency in the xy plane at the elliptic points is $\omega_0(h,I_1,I_2)|_{p^{\pm}_{ell}(h,I_1,I_2)} = \sqrt{2}$, the direction of rotation is preserved for all orbits (so $\omega_0(h,I_1,I_2) \geq 0$), and the frequency monotonically decays as the action of the periodic orbits grows. Denoting by $\omega_{0\,\min}(h) > 0$ the frequency of the two symmetric periodic orbits in the xy plane satisfying $H_{xy}(x,y) = h$, it follows that for this range of energies

$$\omega_{0\,\min}(h) \leq \omega_0(h,I_1,I_2) \leq \sqrt{2}. \tag{6.5}$$

In Figure 6 an example of such a cap shaped energy surface of system (6.1) in the frequency space is shown.

For $h \geq h_{hyp}$ the behavior near the separatrices needs to be presented. Since the frequency in the xy plane is well defined for all orbits except the separatrices, and since $\omega_0(x,y) \to 0$ as the separatrix is approached, defining $\omega_0(0,0) = 0$ makes $\omega_0(x,y)$ a continuous (nondifferentiable) function of the xy energy level. (This observation is used extensively in the frequency map plots; see [28].) Hence, for $H_0 = h \geq h_{hyp}$, an energy surface in the frequency space has an annular cap component which meets at the (singular) circle $\omega_0 = 0$ a central cap;

[12] The existence of such a torus may be formulated as the existence of a transverse intersection of some finite dimensional manifolds. Hence, using the transversality theorem, one proves that it exists for a C^1-open set of integrable Hamiltonians, which we take hereafter as the definition of persistence.

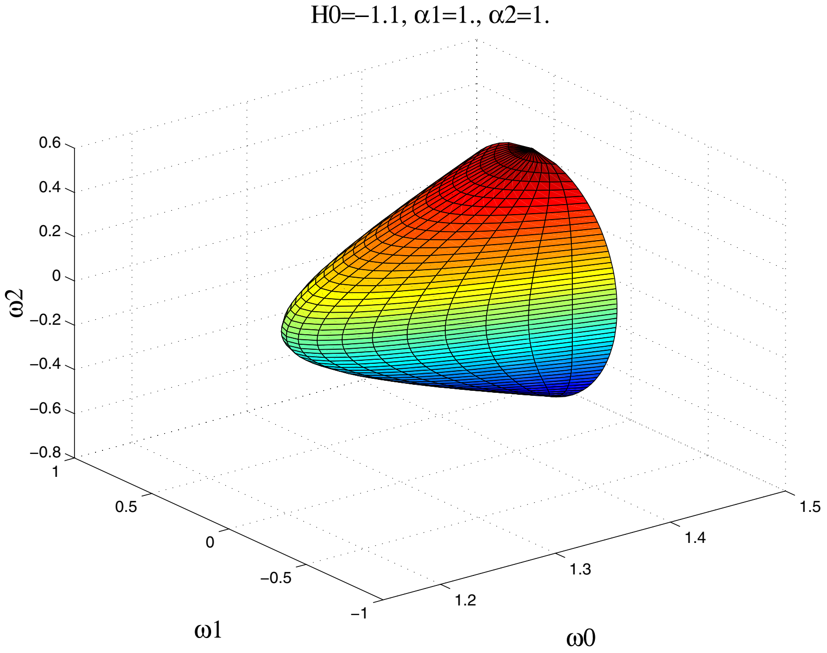


Figure 6. *Energy surface of an a priori unstable system in the frequency space for* $h_{ell} < H_0 < h_{hyp}$.

see Figure 7. The annular cap corresponds to the two sets of tori for which the motion is restricted to one of the wells in the xy plane, whereas the central cap corresponds to a single family of 3-tori for which the motion in the xy plane surrounds both wells of the potential. Recall that strong resonances are created when the energy surface intersects one of the $\omega_i = 0$ planes. However, here, the surface approaches the plane $\omega_0 = 0$ singularly, and the normally hyperbolic torus is not resonant in the θ_0 direction.

6.3. Qualitative behavior of the near-integrable system (U). Using the plots of the EMBD we may read off all possible sources of instabilities for near-integrable n DOF systems with unperturbed Hamiltonian of the form (6.3). Here we need to combine several effects:

- Instabilities associated with the regular resonance web. Such instabilities may appear near any point in the EMBD.
- Instabilities associated with splitting of the separatrices (as in 1.5 DOF systems). Such instabilities may appear for any $h > h_{hyp}$ in an ε neighborhood of the surface $p^0_{hyp}(h, \cdot)$.
- Instabilities associated with the existence of families of separatrices on the same energy surface (as in Arnold's conjecture for the existence of whiskered transition chain). For $n \geq 3$, these appear for any $h > h_{hyp}$ near the surface $p^0_{hyp}(h, \cdot)$.
- Instabilities associated with strongly resonant normally hyperbolic tori. For $k < n-1$, the k-resonant normally hyperbolic tori appear for all $h > h_{hyp}$, and their effect must be included in the above mentioned transition chain.
- Instabilities associated with the topological bifurcation of the energy surface near $h = h_{hyp}$. There, $p^0_{hyp}(\cdot)$ has an $n-1$ fold point and the normally hyperbolic torus $p^0_{hyp}(h_{hyp}, \cdot)$ is a torus of fixed points. Hence, these are the instabilities associated

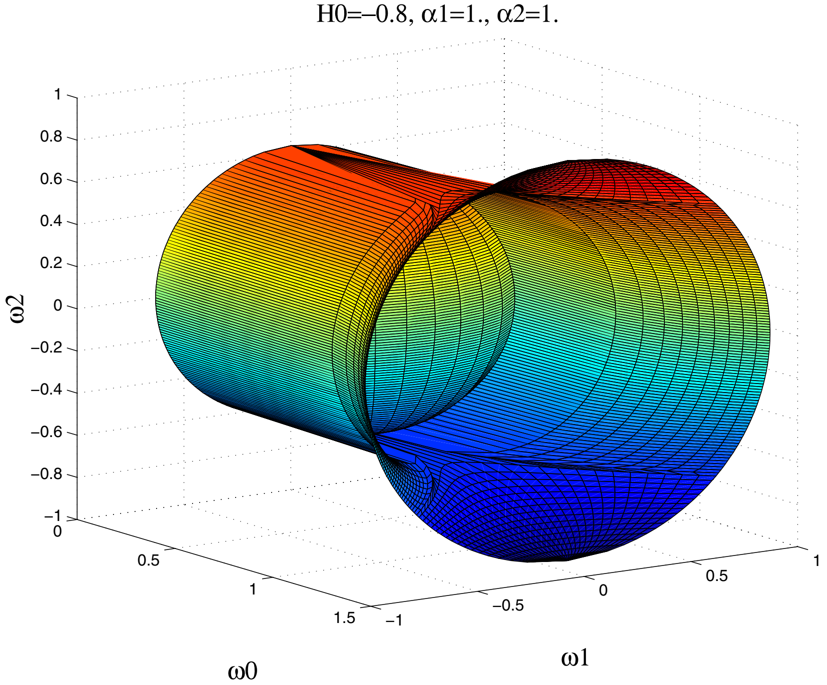


Figure 7. *Energy surface of an a priori unstable system in the frequency space for* $H_0 > h_{hyp}$.

with perturbations of a normally hyperbolic torus of fixed points in the nondegenerate case.

While the analysis of each of the above items is not yet well understood, we propose that the inclusion of rough lower bounds on the instability associated with each of the above phenomena supplies nontrivial information on the system. In Figure 8 we plot on a 2D slice of the EMBD and $O(\varepsilon)$ band around the separatrix level sets (the light shaded region), and indicate an $O(\varepsilon)$ slab of energies to which the perturbed motion is restricted near a hyperbolic resonance (the dark shaded strip). The geometry near the hyperbolic resonant tori immediately presents itself as a source for larger instabilities than the nonresonant terms. The analysis of this case for $n = 2$ has been developed; see [23] and references therein. Here we see that in the 3 DOF context the hyperbolic resonant 2-tori, $p_{hyp}(h, -\alpha_1, I_2(h))$ and $p_{hyp}(h, I_1(h), -\alpha_2)$, belong to the circle of equi-energy normally hyperbolic 2-tori, $p_{hyp}(h, \cdot)$; hence, one is lead to the study of whiskered transition chains with resonant gaps (see [9] and references therein). The subject of transition chains of whiskers in a priori unstable systems has received much attention in recent years (see, e.g., [8, 49, 50] and references therein). Furthermore, as $(I_1, I_2) \to (-\alpha_1, -\alpha_2)$ we approach a double resonant hyperbolic torus—in this case the 3D figure corresponds to a revolution of the EMBD in Figure 8 around the starred line, with the strong resonant planes intersecting as in Figure 4. Here the radius (in I) of the circle $p_{hyp}(h, \cdot)$ scales as $\sqrt{\varepsilon}$ (see Figure 8); hence the transition chain created near such a double resonant hyperbolic torus cannot create large instabilities. If the terms in I are indefinite near such a torus, this situation may change (though our preliminary numerical

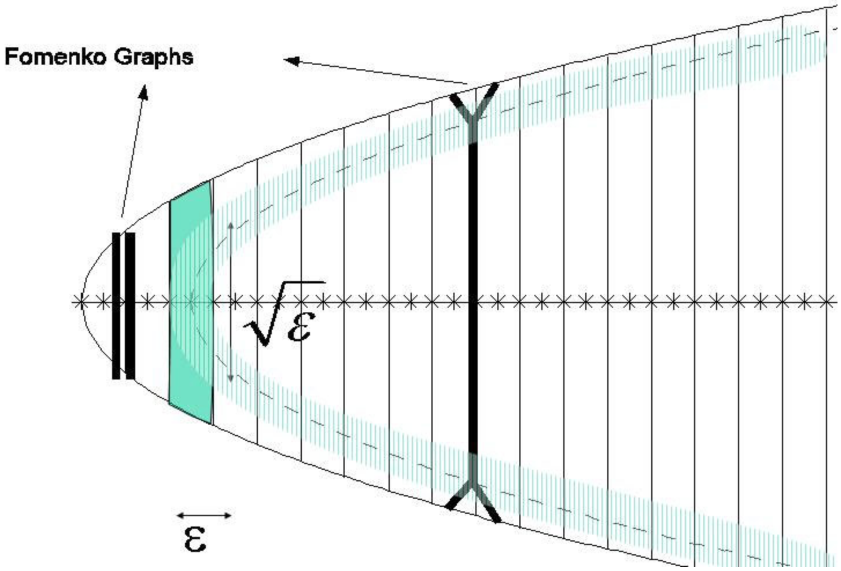


Figure 8. *2D slice of the EMBD of the perturbed motion. Shaded strip: Region of allowed motion for perturbed orbits near hyperbolic resonance. Dashed region: Homoclinic chaos region.*

simulations appear to indicate that even then the instability induced by the separatrices is not significantly enhanced [38]).

Finally, we note that the doubly resonant elliptic tori $p_{ell}^{\pm}(h, -\alpha_1, -\alpha_2)$ reside on two small separate components of the energy surface and hence cannot induce large instabilities, and that other cases corresponding to unbounded energy surfaces may be classified similarly.

7. Bifurcating systems. For n DOF systems with $n \geq 3$ the appearance of parabolic resonant tori is persistent (see [38]); hence their study is both mathematically fascinating and physically relevant. Combining our understanding of the stable and unstable systems, we can now study H_{bif}:

$$H_{bif}(x, y, I_1, I_2) = \frac{y^2}{2} - \frac{x^2}{2}I_1 + \frac{x^4}{4} + \left(\mu_1 + \frac{1}{2}\right)\frac{I_1^2}{2} + \frac{I_2^2}{2} + \alpha_2 I_2 + \alpha_3 I_1 I_2. \tag{7.1}$$

The phase space structure of the Hamiltonian H_{bif} for any fixed I is obvious; for $I_1 > 0$ it is given by the product of a figure eight motion in the xy plane and a family of 2-tori in the (θ, I) space as in H_{ust}, whereas for $I_1 < 0$ it corresponds to an elliptic motion around the origin in the xy plane and a family of 2-tori in the (θ, I) space as in H_{st}. At $I_1 = x = y = 0$ the system has a family of normally parabolic 2-tori. To understand the precise structure of each energy surface, we again construct the EMBD and the corresponding branched surfaces and then present the interesting energy surfaces in the frequency space.

The symmetric form of (7.1) implies that at $I_1 = 0$ we have a pitchfork bifurcation in the xy plane, whereas for a generic asymmetric integrable bifurcating Hamiltonian one should consider instead a saddle-center bifurcation at $I_1 = 0$. The EMBD and branched surfaces

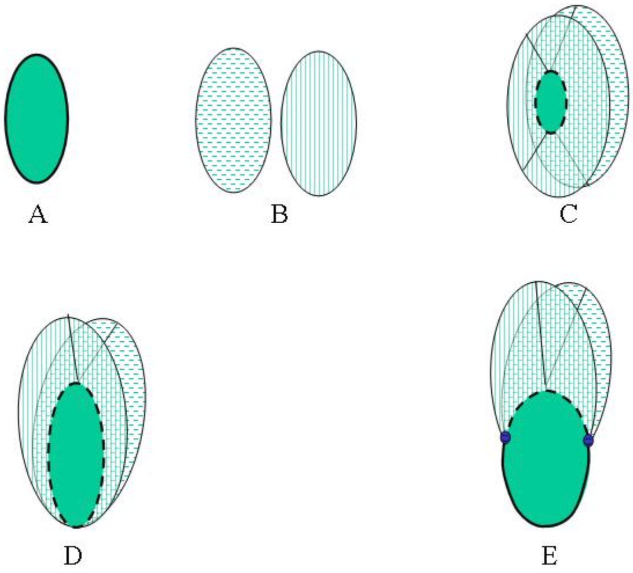


Figure 9. *Generalized Fomenko graphs: The branched surfaces.*

analysis of such systems is analogous to the one presented here. The phenomena of parabolic resonances (PR) in the asymmetric case has not been investigated yet.

7.1. Energy surfaces in the energy-momentum space (B). Recall that the boundary of the allowed region of motion is composed of the singularity surfaces corresponding to lower dimensional normally elliptic tori. For (7.1), these are given by the normally elliptic tori at $\{(x,y)=(\pm\sqrt{I_1},0);\ I_1>0\}$,

$$p^{\pm}_{ell}(h,I_1,I_2)=\left\{(h,I_1,I_2)|\ \mu_1\frac{I_1^2}{2}+\frac{I_2^2}{2}+\alpha_2 I_2+\alpha_3 I_1 I_2=h;\ h\geq h^{+}_{\min},I_1>0\right\}, \tag{7.2}$$

and the elliptic tori at $\{(x,y)=(0,0);\ I_1<0\}$,

$$p^{0}_{ell}(h,I_1,I_2)=\left\{(h,I_1,I_2)\ |\ \left(\mu_1+\frac{1}{2}\right)\frac{I_1^2}{2}+\frac{I_2^2}{2}+\alpha_2 I_2+\alpha_3 I_1 I_2=h;\ h\geq h^{0}_{\min},I_1<0\right\}. \tag{7.3}$$

Another surface of singularity on which the hyperbolic tori $\{(x,y)=(0,0);\ I_1>0\}$ and their separatrices live is given by

$$p^{0}_{hyp}(h,I_1,I_2)=\left\{(h,I_1,I_2)\ |\ \left(\mu_1+\frac{1}{2}\right)\frac{I_1^2}{2}+\frac{I_2^2}{2}+\alpha_2 I_2+\alpha_3 I_1 I_2=h;\ h\geq h^{0}_{\min},I_1>0\right\}. \tag{7.4}$$

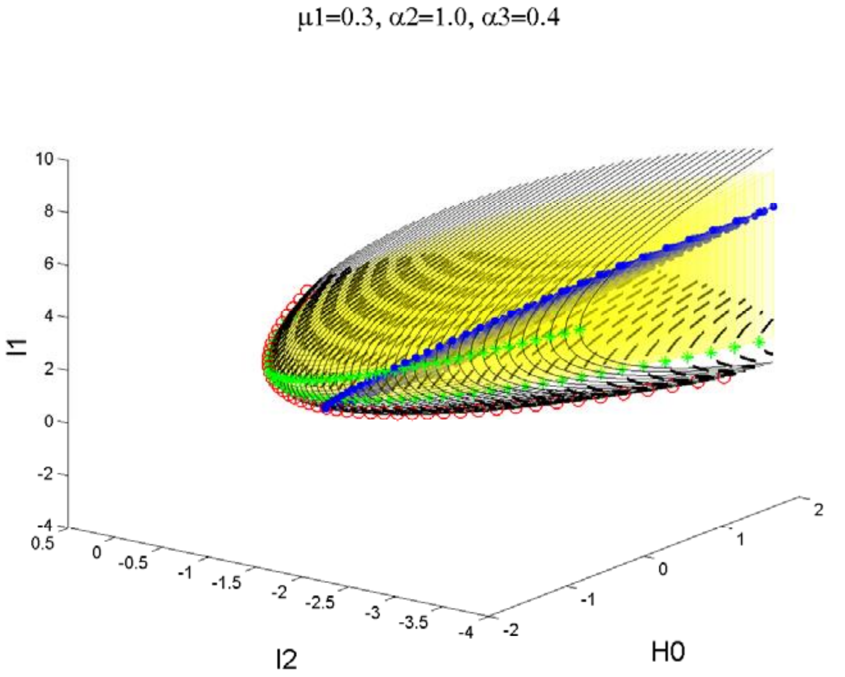


Figure 10. *EMBD for the bifurcating Hamiltonian, H_{bif}.*

Expressing the Hamiltonian on these surfaces in a quadratic form,

$$H_{bif}(0,0,I_1,I_2) = \frac{1}{2}\left(\mu_1 + \frac{1}{2} - \alpha_3^2\right)\left(I_1 - \frac{\alpha_2\alpha_3}{\mu_1 + \frac{1}{2} - \alpha_3^2}\right)^2 + \frac{1}{2}\left(I_2 + \alpha_2 + \alpha_3 I_1\right)^2 - \frac{1}{2}\alpha_2^2 - \frac{1}{2}\frac{(\alpha_2\alpha_3)^2}{\mu_1 + \frac{1}{2} - \alpha_3^2}, \tag{7.5}$$

$$H_{bif}(\pm\sqrt{I_1},0,I_1,I_2) = \frac{1}{2}(\mu_1 - \alpha_3^2)\left(I_1 - \frac{\alpha_2\alpha_3}{\mu_1 - \alpha_3^2}\right)^2 + \frac{1}{2}\left(I_2 + \alpha_2 + \alpha_3 I_1\right)^2 - \frac{1}{2}\alpha_2^2 - \frac{1}{2}\frac{(\alpha_2\alpha_3)^2}{\mu_1 - \alpha_3^2}, \tag{7.6}$$

shows that the sign of $\mu_1 - \alpha_3^2$ determines whether the energy surfaces are bounded in I. Here, for simplicity, we present a bounded case:

$$\mu_1 - \alpha_3^2 > 0, \qquad \alpha_2 > 0, \qquad 0 < \alpha_3 < 1. \tag{7.7}$$

Other cases change some of the inequalities below, leading to a different EMBD and may be similarly analyzed; see [36, 35, 37, 38], where we considered mainly unbounded models. Here, (7.5) and (7.6) define paraboloids, and their intersections with the plane of constant energy define ellipses (see Figures 10, 11, 12, 13, 14, 15, 16).

The minimal energies for which these paraboloids are defined are

$$h^+_{\min} = -\frac{1}{2}\alpha_2^2 - \frac{1}{2}\frac{(\alpha_2\alpha_3)^2}{\mu_1 - \alpha_3^2} < h^0_{\min} = -\frac{1}{2}\alpha_2^2 - \frac{1}{2}\frac{(\alpha_2\alpha_3)^2}{\mu_1 + \frac{1}{2} - \alpha_3^2}, \tag{7.8}$$

with the corresponding minimizing actions

$$\begin{aligned}(I_1, I_2)_{\min 0} &= \left(\frac{\alpha_2\alpha_3}{\mu_1 + \frac{1}{2} - \alpha_3^2}, -\frac{\alpha_2\left(\mu_1 + \frac{1}{2}\right)}{\mu_1 + \frac{1}{2} - \alpha_3^2}\right),\\ (I_1, I_2)_{\min +} &= \left(\frac{\alpha_2\alpha_3}{\mu_1 - \alpha_3^2}, -\frac{\alpha_2\mu_1}{\mu_1 - \alpha_3^2}\right).\end{aligned}$$

Finally, notice that the surface $I_1 = 0$ cuts the paraboloids $p^{\pm}_{ell}(h, I_1, I_2)$ and $p^0_{ell,hyp}(h, I_1, I_2)$ along a parabola which corresponds to a family of normally parabolic 2-tori,

$$p^0_{par}(h, I_1, I_2) = \left\{(h, I_1, I_2) \mid I_1 = 0,\ \ \frac{I_2^2}{2} + \alpha_2 I_2 = h,\ h \geq h^p_{\min}\right\}, \tag{7.9}$$

where

$$h^p_{\min} = -\frac{1}{2}\alpha_2^2 > h^0_{\min}.$$

From these computations we can already conclude that for the chosen set of parameters the energy surface's structure is represented by the branched surfaces shown in Figures 11B–E. Before describing the properties of these surfaces in detail, we examine the appearance of strong resonances so that the topological bifurcations and the appearance of resonant tori may be explicitly related.

Since

$$\omega_2(I_1, I_2) = \frac{\partial H_{bif}(x, y, I_1, I_2)}{\partial I_2} = I_2 + \alpha_2 + \alpha_3 I_1, \tag{7.10}$$

resonances in θ_2 (i.e., resonances in the direction of I_2) are given by the intersection of the domain of allowed motion with the plane

$$I_{2res} = -\alpha_2 - \alpha_3 I_1. \tag{7.11}$$

In particular, resonant lower dimensional tori appear when this plane intersects the paraboloids $p^0_{hyp}(h, \cdot)$, $p^0_{ell}(h, \cdot)$, $p^{\pm}_{ell}(h, \cdot)$, and $p^0_{par}(h, \cdot)$.

Due to the cross term $\frac{x^2}{2}I_1$ in H_{bif} (see (7.1)) we cannot get such a simple and explicit expression for the θ_1-resonant tori surface $\omega_1(I_1, I_2) = 0$. However, the intersection of this surface with the singularity surfaces may be easily found; the hyperbolic (resp., elliptic) lower dimensional tori $p^0_{hyp}(h, \cdot)$ (resp., $p^0_{ell}(h, \cdot)$) are resonant when $\omega_1^0 = 0$, where

$$\omega_1^0 = \frac{\partial H_{bif}(0, 0, I_1, I_2)}{\partial I_1} = \left(\mu_1 + \frac{1}{2}\right) I_1 + \alpha_3 I_2, \tag{7.12}$$

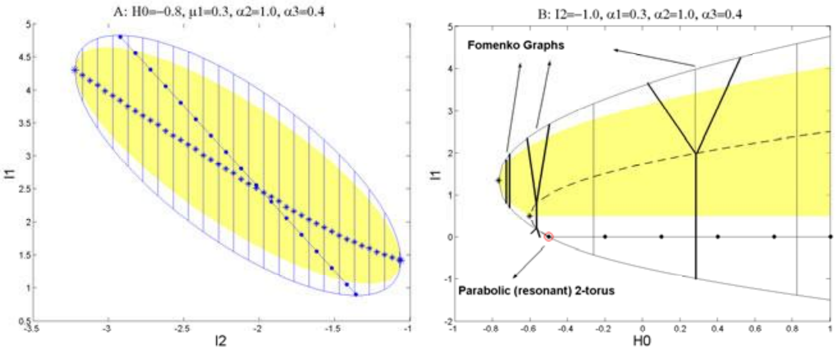


Figure 11. *2D slices of the EMBD of the bifurcating system, H_{bif}.* A: *An energy surface in the energy range $h^{\pm}_{\min} < H_0 < h^0_{\min}$.* B: *An interval of energy values for a fixed value of $I_2 = -\alpha_2$ and all three possible types of (schematic) Fomenko graphs for H_{bif}.*

so resonance occurs exactly at the fold of the singularity surface in the I_1 direction. Hence, the intersection of the plane

$$I_{1res0} = -\frac{\alpha_3}{\mu_1 + \frac{1}{2}} I_2$$

with the paraboloid $p^0_{hyp}(h,\cdot) \cup p^0_{par}(h,\cdot) \cup p^0_{ell}(h,\cdot)$ corresponds to the family of lower dimensional tori that are resonant in the I_1-direction (the green starred curves in Figure 10), $p^0_{res-1}(h, I)$. These tori are normally hyperbolic for $I_{1res} > 0$, normally elliptic for $I_{1res} < 0$, and, at $H_{bif} = h_{par-res1} = 0$, normally parabolic (then $I_{1res} = I_2 = 0$). Similarly, the elliptic lower dimensional tori at $(x, y) = (\pm\sqrt{I_1}, 0)$, $p^{\pm}_{ell}$, are θ_1-resonant when $\omega^{\pm}_1 = 0$, where

$$\omega^{\pm}_1 = \frac{\partial H_{bif}(\pm\sqrt{I_1}, 0, I_1, I_2)}{\partial I_1} = \mu_1 I_1 + \alpha_3 I_2.$$

Hence, the intersection of the plane

$$I_{1res\pm} = -\frac{\alpha_3}{\mu_1} I_2$$

with the paraboloids $p^{\pm}_{ell}(h, I_1, I_2)$ corresponds to these two families of normally elliptic lower dimensional tori that are resonant in the I_1-direction, $p^{\pm}_{res-1}(h, I)$ (denoted by green starred curves in Figure 10 as well). The manifold of 3-tori, which are strongly resonant in θ_1, $p_{res-1}(h, I_1, I_2)$, intersects the paraboloids $p^0_{hyp}(h,\cdot) \cup p^0_{par}(h,\cdot) \cup p^0_{ell}(h,\cdot)$ and $p^{\pm}_{ell}(h, I_1, I_2)$ along the families $p^0_{res-1}(h, I)$ and $p^{\pm}_{res-1}(h, I)$.

Clearly, from the form of (7.1), strong resonance in the xy plane (namely, the normal frequency $\Omega = \omega_0(h, I_1, I_2) = 0$) may occur only at the parabolic tori $p^0_{par}(h, 0, I_2) = p^0_{hyp}(h, 0, I_2) = p^{\pm}_{ell}(h, 0, I_2)$, where $\omega^{\pm}_0 = \sqrt{2I_1}$ and $\omega^0_0 = \sqrt{-I_1}$ vanish. Notice that here, at $I_1 = 0$, the natural frequency of the lower dimensional torus does vanish, as opposed to the formal definition of vanishing ω^0_0 which we had introduced for $I_1 > 0$.

In Figure 10 a 3D EMBD of the system (7.1) (Hamiltonian H_{bif}) is presented for typical parameter values ($\mu_1 = 0.3$, $\alpha_2 = 1$, $\alpha_3 = 0.4$ in all the EMBD plots, and the corresponding

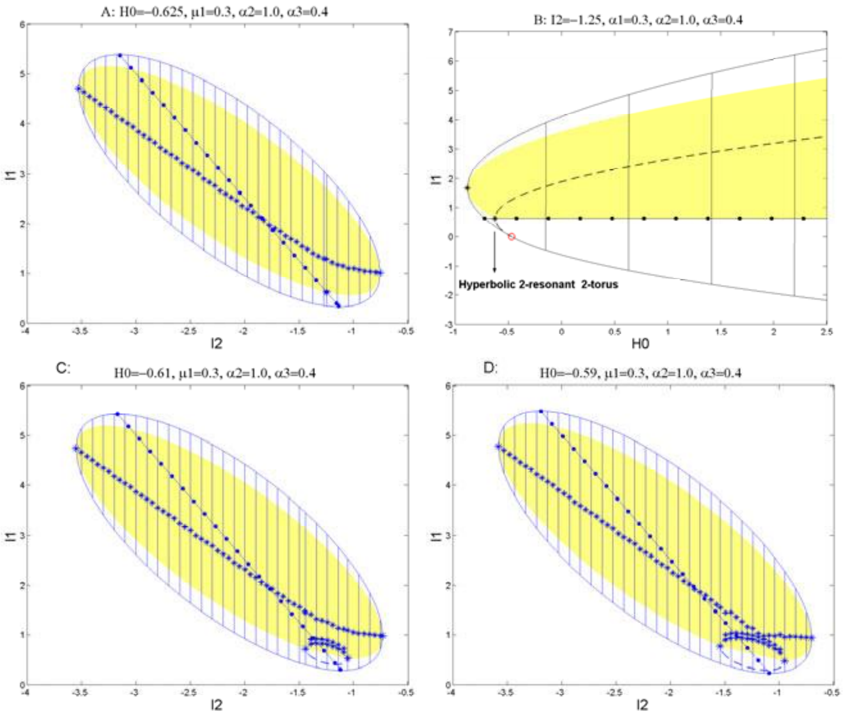


Figure 12. *2D slices of EMBD showing the energy range $h^0_{\min} \leq H_0 < h^0_p$.* A: *An energy surface corresponding to the energy value $H_0 = h^0_{\min}$, showing the hyperbolic bifurcation point.* B: *A range of energy values which includes the bifurcation value at which a hyperbolic double resonance occurs for a fixed value of $I_2 = I_{2\,\min\,0}$.* C, D: *Energy surfaces containing the singular ellipse (dashed line) corresponding to hyperbolic 2-tori and their separatrices.*

energy bifurcation values are $h^+_{\min} \approx -1.0714$, $h^0_{\min} = -0.625$, $h^p_{\min} = -0.5$, and $h_{par-res1} = 0$). The plot includes a limited set of I_2 values to allow a glimpse into its complicated inner structure: the manifold of solid curves corresponds to the singularity manifold of elliptic 2-tori, $p^{\pm}_{ell}(h, I_1, I_2)$ and $p^0_{ell}(h, I_1, I_2)$, the manifold of dashed curves to hyperbolic 2-tori, $p^0_{hyp}(h, I_1, I_2)$, and the curve of red circles to parabolic 2-tori, $p^0_{par}(h, I_1, I_2)$; the strongest resonance in the I_2-direction ($\dot{\theta}_2 = 0$) is denoted by a blue surface of dotted lines and the 2-tori with the strongest resonance in the I_1-direction ($\langle\dot{\theta}_1\rangle_{xy} = 0$) by green starred curves. The yellow volume (or shaded regions in the 2D EMBD) corresponds to regular 3-tori on which $\dot{\theta}_1$ changes sign (back-flow). The surface of 3-tori which have strong resonance in the I_1-direction is contained in this region.

Taking the 2D slices $H_0 = h$ of Figure 10 for increasing h values, we describe below the structure of the corresponding branched surfaces in Figures 9B–E. The intersections of the strong resonance surfaces $\dot{\theta}_2 = 0$ and $\omega_1(I_1, I_2) = 0$ with these 2D slices are denoted by a dotted line and a starred curve (which is calculated numerically), respectively.

For energies in the range $H_0 = h$, $h^+_{\min} \leq h < h^0_{\min}$, the energy surfaces are composed of two separate components corresponding to oscillations in each of the potential wells (as in the low energy a priori unstable case). The energy surface in the EMBD is an ellipse, so any 2D section of Figure 10 for this range of energies is similar to the EMBD slices of system H_{st} presented in Figure 3; see Figure 11A. The Fomenko graph for any 1D section of this ellipse is simply two segments; see Figure 11B. Equivalently, the branched surfaces are two identical discs as in Figure 9B, and the strong resonance surfaces $\omega_1(I_1, I_2) = \omega_2(I_1, I_2) = 0$ intersect the energy surface transversely, as shown by the starred and dotted lines in Figure 11A and schematically in Figure 14A. In particular, since the two resonant in the I_1 direction lower dimensional elliptic tori, $p^{\pm}_{res-1}(h, I_1, I_2)$, are separated along the circle $p^{\pm}_{ell}(h, I_1, I_2)$ by the two resonant in the I_2 direction lower dimensional elliptic tori, $p^{\pm}_{res-2}(h, I_1, I_2)$, it follows that the curves $p_{res-1}(h, I_1, I_2)$ and $p_{res-2}(h, I_1, I_2)$ of resonant 3-tori must intersect *at least* once in a double resonant 3-torus as shown schematically in Figure 14A; indeed, a transverse intersection of the strongest resonance curves is depicted in Figure 11A.

At $H_0 = h = h^0_{\min}$ (Figure 12A) a hyperbolic resonant bifurcation occurs; the singularity surface $p^0_{hyp}(h, I_1, I_2)$ appears in a 2-fold (since $I_{1\min 0} > 0$), creating a torus of fixed points which is normally hyperbolic; see Figure 12B, where the 2D slice $I_2 = I_{2\min 0}$ of Figure 10 is presented, and the newly born 2-resonant hyperbolic torus appears there in the fold of the dashed curve. Figure 12A shows that this torus does not appear at the intersection of the two strong resonance curves $\omega_1 = \omega_2 = 0$ (as in the a priori unstable case and as depicted schematically in Figure 14B). It appears as an isolated star[13] ($\dot{\theta}_1 = \omega^0_1 = 0$) residing on the dotted line ($\dot{\theta}_2 = \omega_2 = 0$). Topologically, at this point the two disks of the branched surfaces meet, so that for $H_0 = h$, $h^0_{\min} < h < h^p_{\min}$, we have, as before, a ring of I values for which two families of 3-tori, corresponding to oscillation in the wells, coexist and a central disk of I values for which only one family of 3-tori, corresponding to motion around the two wells, exists. The boundary between these regions is an ellipse of I values, corresponding to normally hyperbolic 2-tori; see Figures 12C,D and the corresponding Figures 9C and 5A.

Using the computation of $p^0_{res-1,2}(h, I)$, the *minimal number* of intersections of the strong resonance curves $p_{res-1}(h, I_1, I_2)$ and $p_{res-2}(h, I_1, I_2)$ in the interior disk is found to be one, as shown in Figure 14C. The 2D sections of Figure 10 for these ranges of energies demonstrate that for the bifurcating system (7.1) additional intersections appear[14] (see Figure 12A,C,D). A bifurcation in the iso-energetic strong resonance curve $\omega_1 = 0$ occurs at $H_0 = h = h^0_{\min}$. For $h < h^0_{\min}$ this resonance curve has one component and it is smooth; at the bifurcation point $h = h^0_{\min}$ it splits to two components—a smooth curve of resonant 3-tori with two resonant elliptic 2-tori at its boundary and a resonant hyperbolic 2-torus as a separate component (see Figure 12A). For $h^0_{\min} < h < h^c_{\omega_1}$, the iso-energetic curve $\omega_1 = 0$ has three components: the smooth component of resonant 3-tori with elliptic resonant 2-tori as its boundary and the other two components, which reside above and below the ellipse of hyperbolic 2-tori and meet at the two resonant hyperbolic 2-tori, as seen in Figure 12B. As the energy value increases, the components of $\omega_1 = 0$ approach each other until at the next bifurcation point of this curve, $H_0 = h = h^c_{\omega_1}$, all three components meet again (for the parameters chosen here $h^c_{\omega_1} \approx -0.59$; see Figure 12D), forming, for $h > h^c_{\omega_1}$, one nonsmooth component with cusp points at the resonant hyperbolic 2-tori, as seen in Figure 13A. (So the resonant surface $\omega_1 = 0$ folds in the shape of a nose looking toward the negative h values in the EMBD.)

[13] Look *below* the starred curve, at the boundary of the yellow shaded region.

[14] One can argue that this more complicated scenario is the generic one.

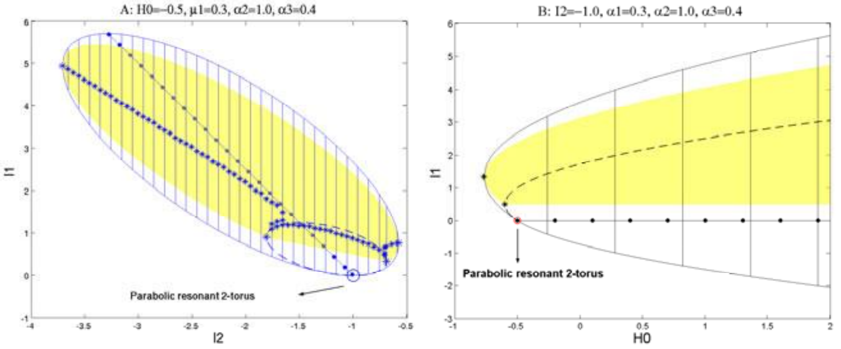


Figure 13. *$2D$ slices of EMBD of the bifurcating system, H_{bif}, showing the parabolic bifurcation point.* A: *An energy surface corresponding to the energy value $H_0 = h^p_{\min}$.* B: *A range of energy values which includes the bifurcation value $H_0 = h^p_{\min}$ for a fixed value of $I_2 = -\alpha_2$.*

At $H_0 = h = h^p_{\min}$ the ellipses $p^0_{hyp}(h^p_{\min}, \cdot)$ and $p^{\pm}_{ell}(h^p_{\min}, \cdot)$ touch the $I_1 = 0$ plane at $I_2 = -\alpha_2$, creating a parabolic torus (see Figure 13A and the schematic representation of Figure 9D). This bifurcation is, again, associated with a fold in the singularity surfaces of the EMBD and therefore with a resonance of a lower dimensional torus (see Theorems 4 and 5 in section 8.2). Indeed, at $H_0 = h^p_{\min}$, the parabola $p^0_{par}(h, I_1, I_2)$ folds in the I_2 direction at $I_2 = -\alpha_2$; hence this parabolic 2-torus is resonant in θ_2:

$$\dot{\theta}_2\Big|_{(0,0,0,-\alpha_2)} = \frac{\partial H_{bif}(0,0,0,I_2)}{\partial I_2}\Big|_{I_2=-\alpha_2} = 0.$$

Furthermore, parabolicity of this torus implies that the torus is strongly resonant in the xy plane; namely, $\omega^0_0 = 0$, so at $h = h^p_{\min}$ a double resonance occurs at the torus $(x, y, I) = (0, 0, 0, -\alpha_2)$. This coincidence of resonances and of topological changes in the energy surfaces is shown schematically in Figures 14C–E. Figures 12 and 13 demonstrate its occurrence for (7.1). The dotted line, representing resonant tori in the I_2-direction, intersects the boundary of the allowed region of motion at the parabolic torus (a circle in the figure).

For $h > h^p_{\min}$, the branched surface is a disk with two flaps emanating from it (see Figure 9E), where the two end points of the flaps correspond to parabolic tori. For Hamiltonian (7.1), the ellipse $H_{bif}(0, 0, I_1, I_2) = h$ defines the boundary of this disc. The ellipse corresponds to hyperbolic tori for $I_1 > 0$ (the dashed part of the inner ellipses in Figure 15) and elliptic tori for $I_1 < 0$ (the lower solid part of the ellipses in Figure 15). The flaps' upper boundaries (upper solid line in Figure 15) are defined by the ellipse $H_{bif}(\pm\sqrt{I_1}, 0, I_1, I_2) = h$ for $I_1 \geq 0$. The two meeting points of the flaps (denoted by circles) with this disk correspond to the two parabolic tori which reside on the energy surface (see (7.9)).

Now, consider the relative location of the resonant in the I_1-direction 2-tori, $p^0_{res-1}(h, I)$, and the parabolic 2-tori, $p^0_{par}(h, I)$, on the ellipse $H_{bif}(0, 0, I_1, I_2) = h$. For h values which are slightly larger than $h^p_{\min}$, this pair of resonant 2-tori lives on the upper part of the ellipse, above the parabolic tori; hence they correspond to normally hyperbolic resonant 2-tori, as shown schematically in Figure 14D. For such values of h the curve $p_{res-1}(h, I_1, I_2)$ intersects only the upper part of the ellipse (the dashed part of the ellipse in Figure 15A, corresponding

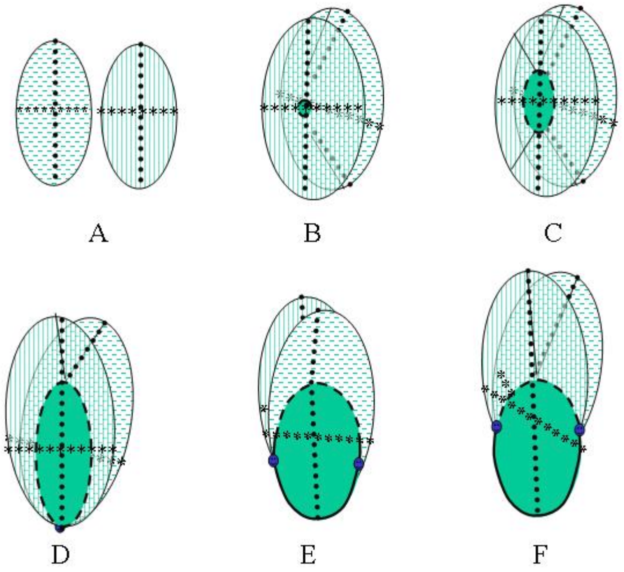


Figure 14. *Energy branched surfaces and strong resonance curves.*

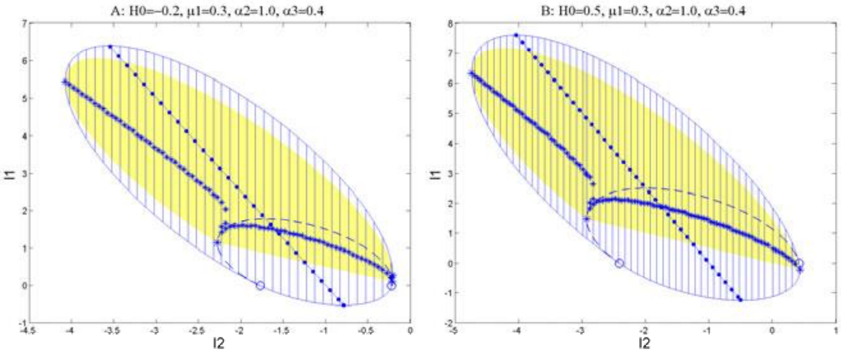


Figure 15. $2D$ *slices of an EMBD of the bifurcating system,* H_{bif}. A: *An energy surface with* $h^p_{\min} < H_0 < h_{par-res1}$. B: *An energy surface with* $H_0 > h_{par-res1}$.

to hyperbolic tori), so there are no resonant in the I_1 direction normally elliptic 2-tori at the origin (i.e., $p^0_{hyp}(h, I_1, I_2) \cap p_{res-1}(h, I_1, I_2) = \varnothing$), as shown in Figure 15A. For energy values greater than $H_0 = h_{par-res1} = 0$ (i.e., for $h > h_{par-res1}$), this situation changes; for $H_0 = h > h_{par-res1}$, the resonant plane $p_{res-1}(h, I_1, I_2)$ intersects each of the curves $p^0_{hyp}(h, \cdot)$ and $p^0_{ell}(h, \cdot)$ at one point (see the schematic Figure 14F and the energy surface in Figure 15B). Namely, one of the resonant hyperbolic lower dimensional tori becomes normally elliptic for $H_0 = h > h_{par-res1}$. Therefore, at the bifurcation value, $H_0 = h_{par-res1} = 0$, the resonant plane in the I_1-direction intersects the ellipse $p^0_{hyp}(h, \cdot) \cup p^0_{par}(h, \cdot) \cup p^0_{ell}(h, \cdot)$ at $I_1 = I_2 = 0$, where a parabolic, resonant in the I_1 direction, lower dimensional torus is created (the schematic Figures 14E,F show the aforementioned intersections before and after

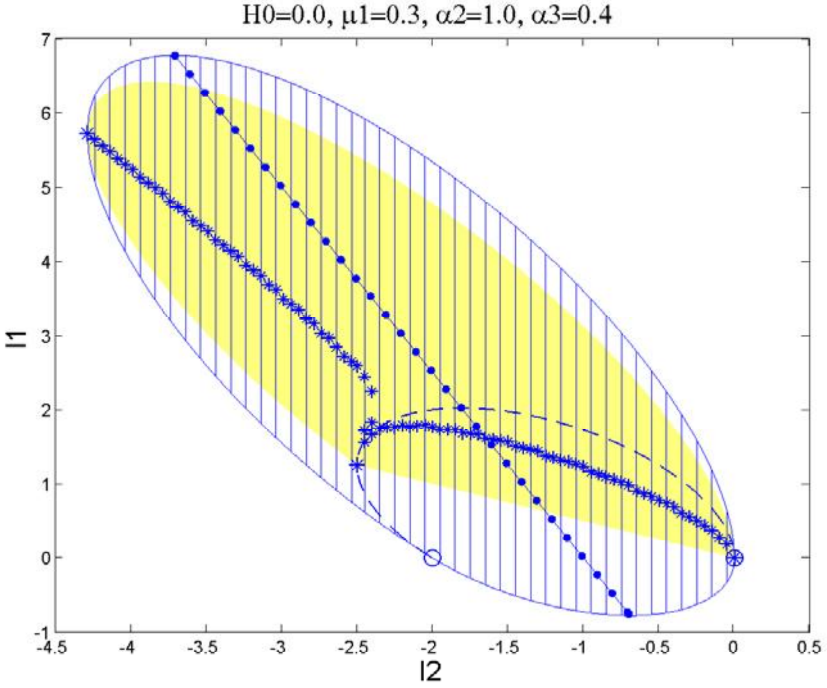


Figure 16. *A 2D slice of EMBD of the bifurcating system, H_{bif}. An energy surface with $H_0 = h_{par-res1}$, containing a strongly resonant in the I_1-direction parabolic 2-torus.*

this bifurcation). Indeed, Figure 16 shows that at $H_0 = h_{par-res1} = 0$ one of the resonant hyperbolic tori changes its stability and becomes parabolic. (The end point of the starred curve, $\omega_1 = 0$, intersects one of the circles denoting a parabolic 2-torus; note that at the bifurcation point $H_0 = h_{par-res1}$, the iso-energetic curve $\omega_1 = 0$ ceases to have two cusp points and thereon has only one cusp point at the remaining resonant hyperbolic 2-torus.) Figure 17A demonstrates that the resonance in the I_1-direction is indeed associated with a fold of the parabola $p^0_{hyp}(h, I_1, 0) \cup p^0_{par}(h, I_1, 0) \cup p^0_{ell}(h, I_1, 0)$ at the origin; the 2D slice of the EMBD at $I_2 = 0$ shows that the circle denoting the parabolic torus and the star denoting the strong resonance in the I_1 direction coincide.

Bifurcation values for the parameters are now easily identified. First, we see that at $\alpha_2 = 0$, $h^+_{\min} = h^0_{\min} = h^p_{\min} = h_{par-res1} = 0$; namely, all the bifurcations mentioned above occur at one energy surface, and a double resonant (torus of fixed points in the 3 DOF case) normally parabolic torus is created, as shown in Figure 17B, where the star ($\omega_1 = 0$), the dotted line ($\omega_2 = 0$), and the circle (a parabolic torus) coincide. Then, the energy surface $H_0 = \alpha_2 = 0$ of system (7.1) shrinks to one 2-resonant normally parabolic 2-torus of fixed points. The existence of a double resonant parabolic torus is a codimension one phenomena for 3 DOF systems and a persistent phenomena in 4 or larger DOF systems [38]. Normally parabolic tori of fixed points are a codimension one phenomenon for any $n \geq 2$ (see [45, 38, 37] for more details).

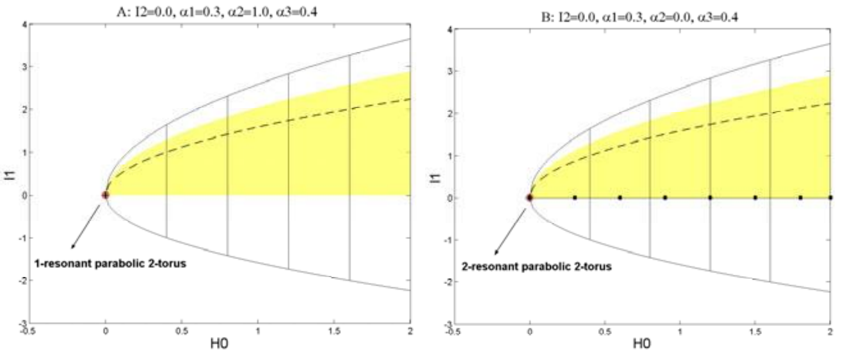


Figure 17. *2D slices of the EMBD at $I_2 = 0$.* A: *Regular parameter values—at $I_1 = H_0 = 0$ a resonance in the I_1 direction occurs.* B: *At the special bifurcation value $\alpha_2 = 0$. Since $\alpha_2 = 0$, at $I_1 = H_0 = 0$ a double resonance in the I_1 and I_2 direction occurs.*

Second, notice that

$$\frac{d^2 H_{bif}(\pm\sqrt{I_1}, 0, I_1, I_2)}{dI_1^2} = \mu_1; \tag{7.13}$$

hence, the fold of the singular surface $p^+_{ell}(h, I_1, I_2)$ in the I_1-direction becomes flatter as $\mu_1 \to 0$ (put differently, the dependence of the frequency $\omega^\pm_1$ on I_1 becomes weaker). It is seen that holding α_3 fixed in this limit changes the character of the energy surfaces from being bounded to being unbounded in I_1, I_2. We will not delve into the analysis of all the different limits which may be taken here; some of these limits are studied in detail in previous works (see [45, 46, 35, 36, 37, 38]). In particular, note that the appearance of flat parabolic resonant tori in such a situation gives rise to strong instabilities (see [38] and [37]).

7.2. The frequency domain plots (B). We plot the energy surfaces (ω^H) of the bifurcating system in the frequency space for typical and bifurcating energy values ($h^+_{\min}, h^0_{\min}, h^p_{\min}$, and $h_{par-res1}$ and h values in between them) with the resonance web plotted on them. We demonstrate that the structure of these webs differs from the structure of webs of a priori stable systems in its nonuniformity and its behavior near the origin.

The simplest type of energy surface component contains only elliptic lower dimensional tori. For $H_0 = h$, $h^+_{\min} < h < h^0_{\min}$, it appears as a smooth codimension one surface with boundaries, as shown in Figure 18, similar to Figure 6. This smooth compact component is a smooth deformation of the disk appearing in the energy-momentum space, (H_0, I_2, I_1). Transverse intersection of a smooth component of ω^H with one of the planes $\omega_j = 0$ corresponds to a strong resonance; for energy surfaces in the range $h^+_{\min} < h < h^0_{\min}$ this occurs only for $j = 1, 2$ (it cannot occur for $j = 0$ since in this energy range $\omega_0 > 0$). In Figure 18 (and in the following figures here) the red starred curve corresponds to the intersection of the energy surface, ω^H, with the resonance plane, $\omega_1 = 0$, and the black thick dotted line denotes the intersection of ω^H with $\omega_2 = 0$. The lower dimensional elliptic resonant tori correspond to the intersections of the surfaces' boundary, which is plotted in thick black, with the $\omega_j = 0$ ($j = 1$ or 2) planes.

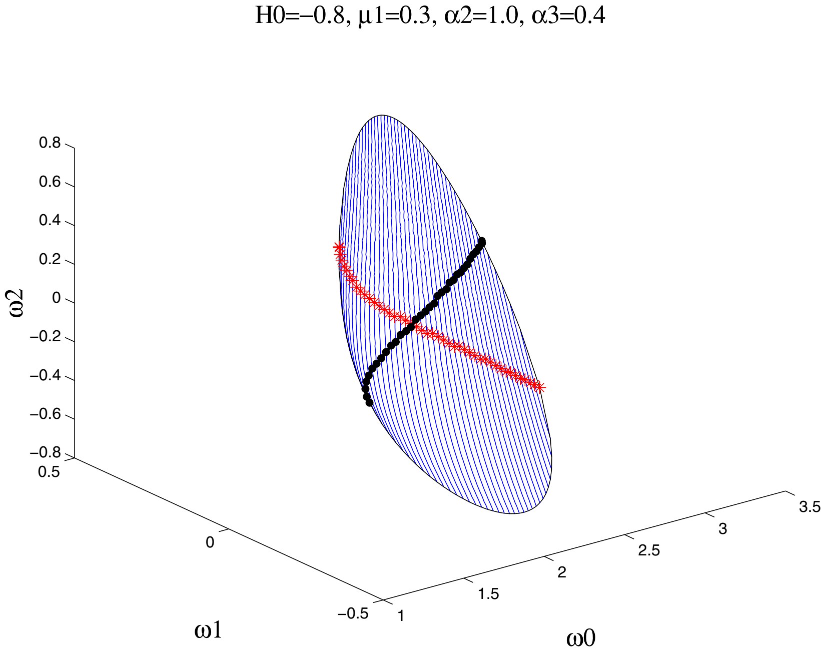


Figure 18. *A typical energy surface corresponding to an elliptic energy value $H_0 = h$ with $h^+_{\min} < h < h^0_{\min}$.*

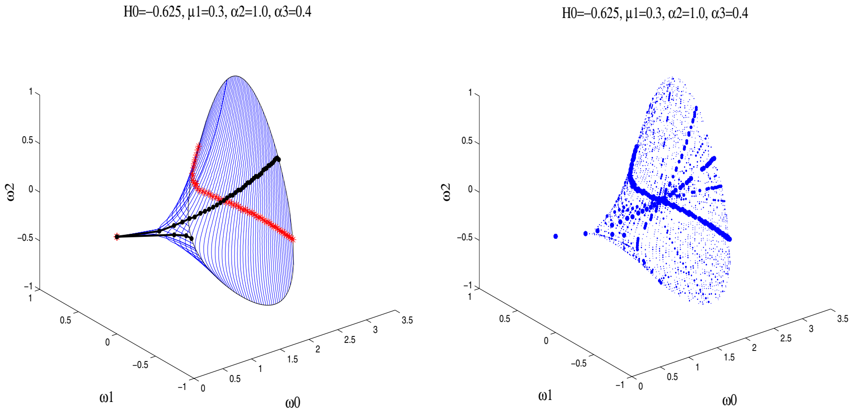


Figure 19. *An energy surface corresponding to the bifurcation value $H_0 = h^0_{\min}$, containing a double resonant hyperbolic torus of fixed points. Left: The frequency map plot. Right: The resonance web on this energy surface for $|k| \leq 21$, where the size of the dots indicates the strength of the resonance.*

The energy value $H_0 = h^0_{\min}$ is a bifurcation value at which one 2-resonant hyperbolic 2-torus (hyperbolic torus of fixed points) appears. It creates a singular cusp point in the energy surface ω^H (see Figure 19), where this energy surface is presented in the three frequency space in the left plot (each blue thin curve corresponds to a fixed value of I_2, the red starred curve to the strong resonance $\omega_1 = 0$, the black dotted line to the strong resonance $\omega_2 = 0$, and the black thin curve to the boundary of ω_H, consisting of 2-tori) and the resonance web on

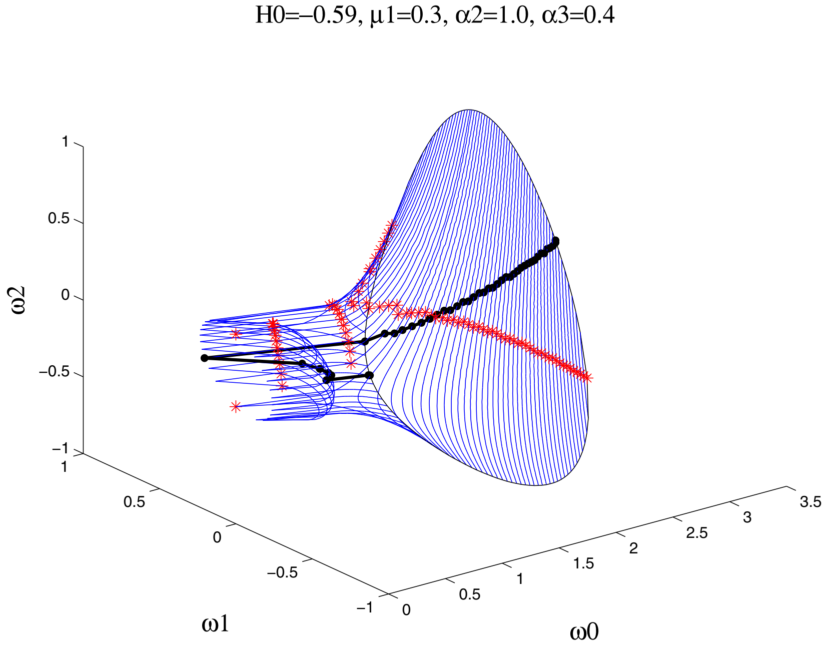


Figure 20. *A typical energy surface in the hyperbolic energy range, corresponding to an energy value $H_0 = h$ with $h^0_{\min} < h < h^p_{\min}$.*

this energy surface is presented in the right plot. The resonance webs presented here are calculated by finding (approximately) the points on the energy surface for which $\langle k, \omega^H \rangle = 0$ for $0 < |k| = |k_1| + |k_2| + |k_3| \leq 21$, where the size of the dots is in inverse relation to $|k|$; i.e., the stronger the resonance the larger the dot indicating it (note that for the weaker resonances the difference in the size of the dots is indistinguishable). The hyperbolic 2-resonant 2-torus in Figure 19 resides in the cusp, far from the other resonance lines and from the main resonance junction, where the strong resonances intersect (this might suggest an additional reason for not observing strong instabilities of the perturbed system near such hyperbolic double resonances [38]).

Energy surfaces with $H_0 = h$, $h^0_{\min} < h < h^p_{\min}$, include an ellipse of hyperbolic tori with their separatrices. As for the a priori unstable case, we find that the energy surface ω^H collides at this singular ellipse with the plane $\omega_0 = 0$ and then bounces back with the same sign of ω_0 (since the direction of motion does not change from the exterior to the interior tori). Using (7.12) and (7.10) we find that the singularity manifold corresponding to the family of hyperbolic 2-tori is given by

$$\begin{aligned} H_{bif}(0,0,I_1,I_2) &= H_{bif}(0,0,\omega_1,\omega_2) \\ &= \frac{1}{2}\frac{\left(\omega_1^0 - \alpha_3\omega_2\right)^2}{\mu_1 + \frac{1}{2} - \alpha_3^2} + \frac{1}{2}\left(\omega_2\right)^2 - \frac{1}{2}\alpha_2^2 - \frac{1}{2}\frac{\left(\alpha_2\alpha_3\right)^2}{\mu_1 + \frac{1}{2} - \alpha_3^2}. \end{aligned} \tag{7.14}$$

For our parameter range it is a tilted ellipse lying in the $\omega_0 = 0$ plane which is centered at the origin; see Figure 20 (similar to Figure 7).

On one side of this singularity manifold each point on the energy surface corresponds to two 3-tori, and on the other side to a single 3-torus as summarized by Figure 9C. Each of the surfaces $\omega_j = 0$ $(j = 1, 2)$ intersects the ellipse at two points, at hyperbolic 1-resonant 2-tori. In [38] we prove that such intersections are persistent. Recall that even though the singular circle is contained in the $\omega_0 = 0$ plane it does not correspond to a double resonance of the lower dimensional torus: $\omega_0 = 0$ at the homoclinic loop, whereas the normal frequency of the hyperbolic torus is imaginary and is nonzero.

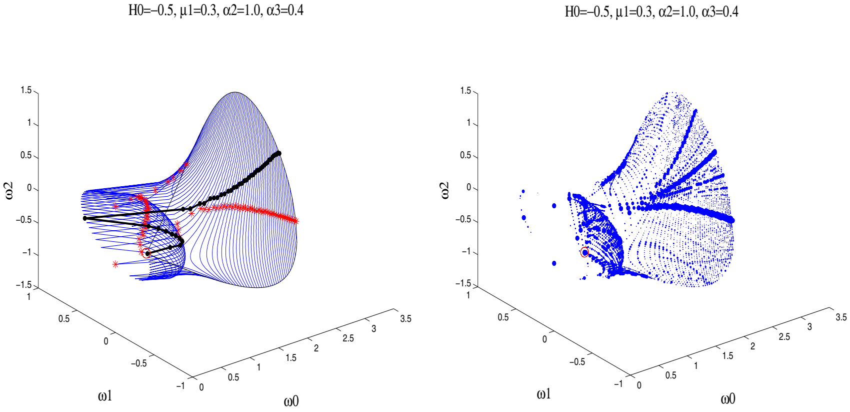


Figure 21. *An energy surface with $H_0 = h^p_{\min}$. Left: The energy surface in the frequency space. Right: The resonance web on this energy surface. Red circle: (Double) resonant (in the I_0- and I_2-directions) parabolic 2-torus.*

At the bifurcation value $H_0 = h^p_{\min}$ a parabolic (resonant in the I_2-direction) torus first appears; see Figure 21. An important observation is that *parabolic tori are a priori resonant*: their normal frequency vanishes. Indeed, let $\omega = (\Omega, \omega^{n-1}) \in \mathbb{R}^n$ denote the n dimensional vector of frequencies, including the normal frequency Ω, and the inner frequencies $(\omega^{n-1} \in \mathbb{R}^{n-1})$ of the $(n-1)$-torus. Parabolicity implies $\Omega = 0$; hence, $k^1 = e^1_n = (1, 0, \dots, 0)$ satisfies the resonance condition $\langle k^1, \omega \rangle = 0$ (indeed, in the resonance web plots a large dot indicating strongest resonance always appears on the parabolic tori; see, e.g., Figure 21). Lower dimensional resonance implies that there exists at least one additional vector of integers, $k^2 = (0, l_{n-1}), l_{n-1} \in \mathbb{Z}^{n-1}$, such that $\langle k^2, \omega \rangle = 0$. Hence, *parabolic lower dimensional resonant tori correspond to junctions in the resonance web with at least one strongest resonance* (indeed, the parabolic torus in Figure 21 is doubly resonant, residing on the junction $\omega_0 = \omega_2 = 0$). In particular, if the parabolic torus appears at the origin, where all resonances intersect, it corresponds to an $(n-1)$-resonant $(n-1)$-torus, namely to a parabolic torus of fixed points. In [38] we prove that such a scenario is persistent in a one parameter family of integrable n DOF Hamiltonian systems with $n \geq 2$.

A typical energy surface in the energy range $H_0 = h$, $h^p_{\min} < h < h_{par-res1} = 0$, is shown in Figure 22 and in the range $H_0 = h > h_{par-res1} = 0$ in Figure 23, where the two parabolic tori are denoted by red circles. Then, the natural frequency in the xy-direction found from

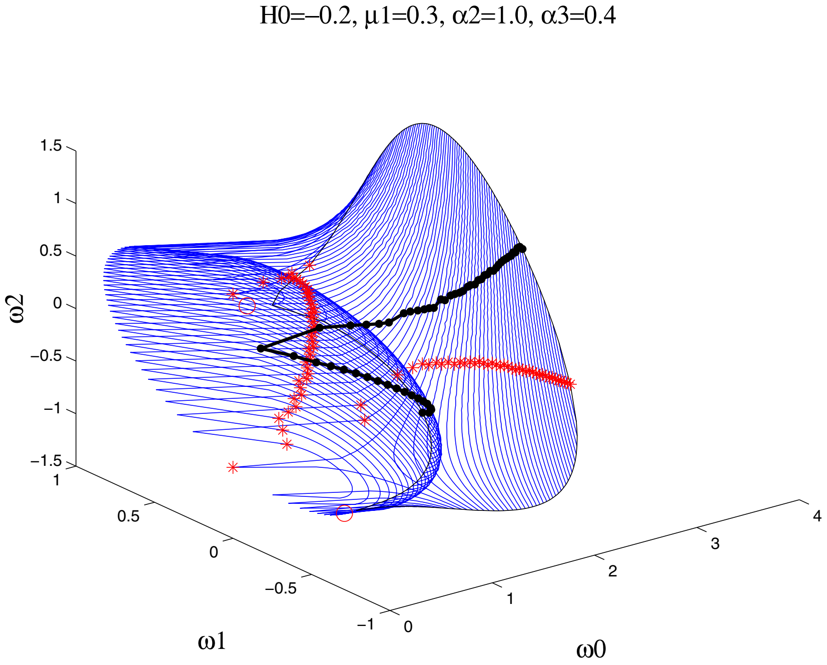


Figure 22. *A typical energy surface with* $h^p_{\min} < H_0 < 0$.

linearization at the origin,

$$\omega_0^0 = \sqrt{-I_1} = \sqrt{\frac{\omega_1^0 - \alpha_3\omega_2}{\mu_1 + \frac{1}{2} - \alpha_3^2} - \alpha_2\alpha_3},$$

shows that the singularity ellipse (7.14) detaches from the $\omega_0 = 0$ plane with a square-root distance. Topologically, the energy surface is well described by the branched surface in Figure 9E.

The colliding surface, at which ω^H is singular (nonsmooth), is clearly of codimension two, and it corresponds to the family of hyperbolic tori which live on the given energy surface. The end points of this collision surface, where the projection singularity heals and the energy surfaces cease to contain hyperbolic tori, correspond to parabolic tori, a codimension three surface, namely, points in Figures 9, 21, 22, 23, and 24 (the parabolic tori are denoted by red circles). At the parabolic lower dimensional tori the ω_0 frequency vanishes. If such an end surface (in the figures, a point) intersects another resonance surface, a parabolic (doubly) resonant torus is born. It is now clear that with additional DOF such an intersection (of the boundary of the collision surface and the resonances on the $\omega_0 = 0$ plane) is generically transverse (see [38] for a proof); hence parabolic resonances (PR) are expected to occur on surfaces corresponding to a range of energies. For the 3 DOF case, since generically the end points (corresponding to the inner frequencies of the parabolic tori) change continuously with the energy values, there exists a set of dense values of energies for which these end points hit resonance surfaces and PR are created. When an end point of a singularity curve belongs to

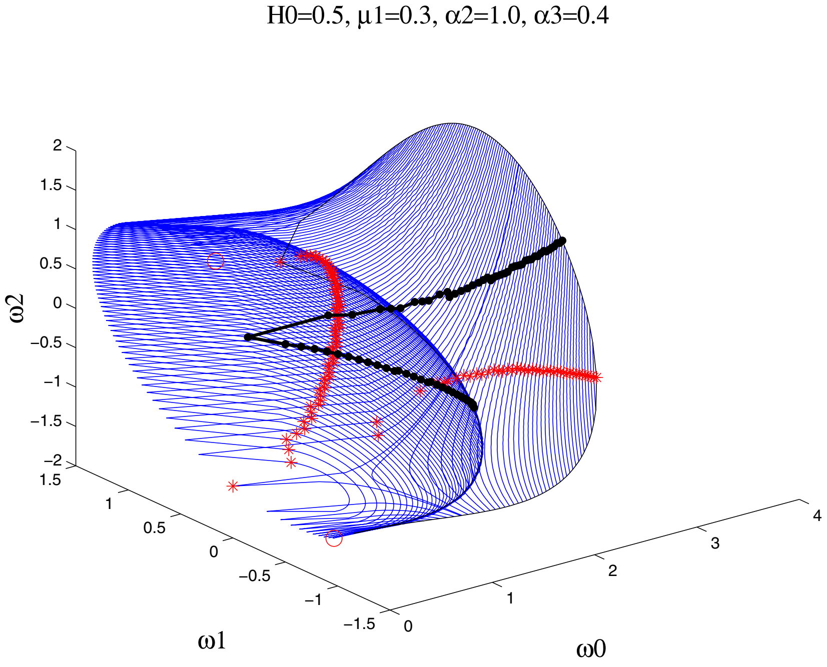


Figure 23. *A typical energy surface in the range $H_0 > h^0_{par-res1} = 0$.*

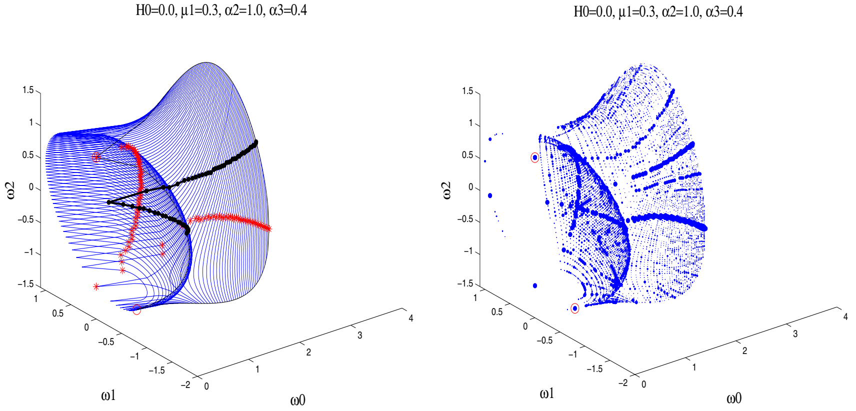


Figure 24. *An energy surface with $H_0 = h_{par-res1} = 0$, containing a strongly resonant in the I_1 direction parabolic 2-torus (hence, a double resonant parabolic torus). Left: This energy surface in the frequency space. Right: The resonance web on this energy surface.*

a strong resonance plane $\omega_j = 0$ ($j = 1$ or 2) it corresponds to a strong *double resonance* of the parabolic lower dimensional torus (see the resonance webs in Figures 21 and 24).

Figure 23 shows an energy surface for positive H_0, where the family of tori encircling the two wells crosses the $\omega_1 = 0$ plane. Figure 24 shows an energy surface with $H_0 = h_{par-res1} = 0$, which contains a resonant in the I_1 direction parabolic torus (hence strongly doubly resonant with $\omega_0 = \omega_1 = 0$) and the resonance web on this energy surface. Setting (in addition) $\alpha_2 = 0$,

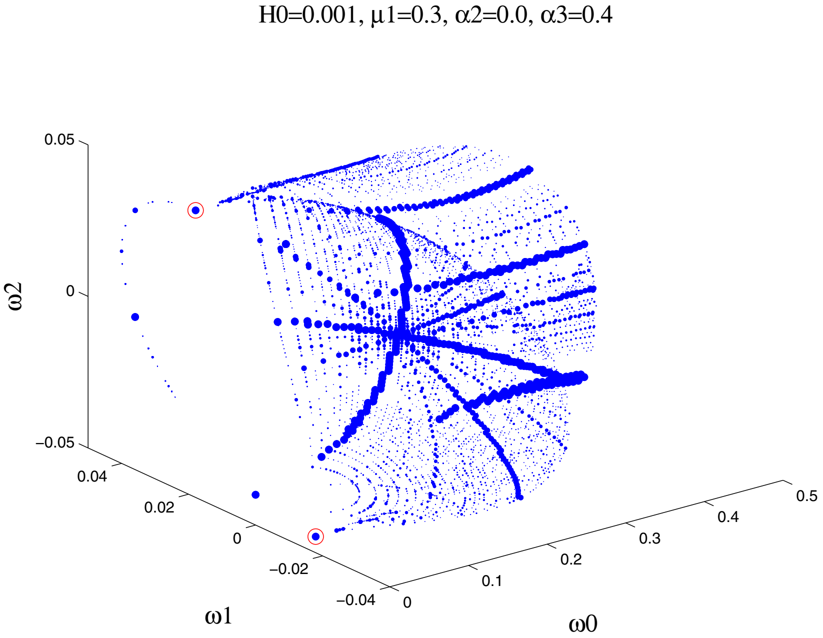


Figure 25. *A resonance web on an energy surface near the (locally KAM degenerate) energy surface with $H_0 = \alpha_2 = 0$.*

the energy surface with $H_0 = 0$ shrinks to a parabolic torus of fixed points at the origin of the frequency space. However, nearby energy surfaces (i.e., energy surfaces with $\alpha_2 = 0$ and a small energy value) have a nondiminishing extent in the frequency space, with resonant parabolic 2-tori residing near the main junction where many strong resonances intersect; see Figure 25 (note the scale of the axis).

Summarizing, we discovered that the presentation in the frequency space of the energy surfaces of Hamiltonians of the form $H_0(x, y, I)$ with $n-1$ dimensional tori that change their stability has the following properties:

- For a range of energies, the energy surface is singular along a codimension two surface belonging to the $\omega_0 = 0$ plane. This singularity surface corresponds to hyperbolic lower dimensional tori and their separatrices. The boundaries of the singularity surface (of codimension three) correspond to parabolic tori.
- Parabolic resonant tori may be recognized as resonance junctions which belong to the boundary of the hyperbolic singular surface. For 3 DOF systems these appear on a dense set of energy values; for $n \geq 4$ these appear for a range of energies.
- While the resonance surfaces still intersect the energy surfaces densely, the *uniformity seems to be lost.*
- A parabolic torus of fixed points appears when the boundary of the singular surface contains the origin. Such a scenario appears for special parameter values (a codimension one phenomena) and on specific energy surfaces.

7.3. Qualitative behavior of the near-integrable system (B). Using the plots of the EMBD we may read off all possible sources of instabilities. Here we need to combine several effects:

- Instabilities associated with the regular resonance web, as in the elliptic case.
- Instabilities associated with the existence of equi-energy family of separatrices and their resonances, as in the unstable case.
- Instabilities associated with resonant parabolic tori; their appearance implies the coexistence of equi-energy families of separatrices and equi-energy families of lower dimensional elliptic tori, meeting at the parabolic tori. The flatness of the singularity manifolds (see (7.13)) affects the extent of the instability.
- Instabilities associated with bifurcations in the structure of the singularity manifolds (manifolds corresponding to lower dimensional tori) of the energy surfaces—namely, the creation of elliptic, hyperbolic, and parabolic lower dimensional tori, all of which are associated with resonant lower dimensional tori.

Once again, the analysis of each of the above items has not been done yet. For the parabolic case we have mainly numerical indications for the behavior of the perturbed orbits; initial steps of a rigorous analysis of instabilities associated with PR are presented in [39] (a longer detailed version is in preparation). The behavior near a nonresonant parabolic torus does not yield instability—the lower dimensional parabolic torus persists [24]—and it appears that the behavior near it is indistinguishable from that appearing near the lower dimensional normally elliptic torus. However, numerical simulations indicate that the behavior near PR is dramatically different; orbits which appear to be chaotic and of a different nature than the homoclinic chaos are abundant. The structure of these perturbed orbits near 1-PR, which appear for a dense set of energy values, is similar to the one observed in the 2 DOF case; see [37, 45]. Further degeneracies make the instabilities more pronounced, see [37, 38] for examples.

One degeneracy we explore here is the existence of a normally parabolic torus of fixed points which is of codimension one ($\alpha_2 = 0$) and corresponds to *a local* violation of the KAM nondegeneracy condition. The induced strong instabilities of a perturbed orbit with initial values near this point are presented in Figures 26 and 27; in Figure 26 the perturbed orbit is projected on the (θ, I) planes, where its complicated structure, while it passes through the successive resonance zones, may be seen; in Figure 27 we show the development of the instabilities in the action variables depending on time. These figures were produced for the perturbed Hamiltonian:

$$H^{\varepsilon}_{bif}(x, y, \theta_1, I_1, \theta_2, I_2; \varepsilon) = \frac{y^2}{2} - \frac{x^2}{2} I_1 + \frac{x^4}{4} + \left(\mu_1 + \frac{1}{2}\right) \frac{I_1^2}{2} + \frac{I_2^2}{2} + \alpha_3 I_1 I_2 + \varepsilon \left(\left(1 - \frac{x^2}{2}\right) \cos(3\theta_1) + \cos(3\theta_2) \right). \tag{7.15}$$

Graphically, in the frequency space, such a scenario happens when the boundary of the singularity surface (here the end points of the singularity lines) passes through the origin, where all the resonance planes intersect. The fact that a parabolic 2-resonant torus resides at this junction point seems to induce strong instabilities in the perturbed system in both

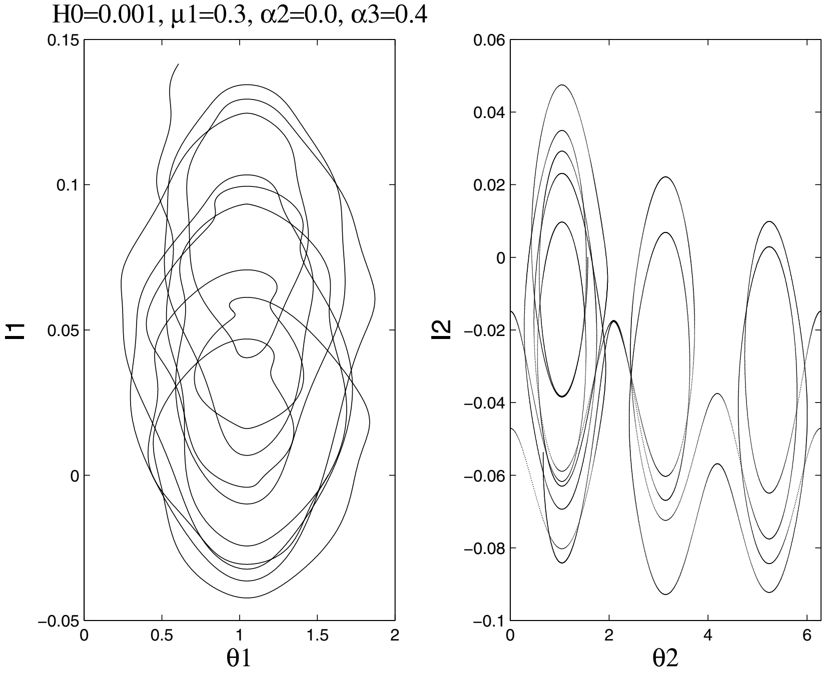


Figure 26. *Instability in the action variables near 2-PR: A perturbed orbit projected on the* (θ, I) *planes, corresponding to the total energy* $H \approx 9.953e-4$*, with initial conditions* $(x, y, \theta_1, I_1, \theta_2, I_2; \varepsilon) = (0.2515, 0, 1.57, 0, 1.57, 0, 1e-3, 1e-3)$*.*

action directions, as seen in Figures 26 and 27. In Figure 28 the corresponding unperturbed energy surface is shown in the (I_2, I_1) plane and in the frequency space (see the corresponding resonance web in Figure 25). The perturbed orbit shown in Figures 26 and 27 approximately covers the whole possible extent of the actions range on this surface. For more details and upper bounds on the maximal instability rate see [37, 38, 39]; in particular, in [39] analytical methods for studying instabilities near 2-PR are suggested.

Note that in 4 or more DOF systems the existence of a double resonant parabolic torus is persistent without dependence of the system on external parameters, and the local violation of the KAM iso-energetic nondegeneracy condition is avoided. Numerical simulations suggest that near such double PR the instabilities and the orbit structure are similar to the ones appearing in the locally degenerate 3 DOF system (with $\alpha_2 = 0$).

8. Bifurcations in the energy-momentum diagrams. Here we formulate the observed relations between bifurcations in the EMBD and the appearance of lower dimensional resonances precisely. First we prove that extrema of the nonparabolic singularity surfaces in the EMBD occur iff the corresponding tori are resonant. Then we prove that if on a given energy surface there exists an $(n-1)$-nonparabolic torus which is nondegenerately strongly $(n-1)$-resonant (see definitions below), then the topology of the energy surfaces changes at this value of the energy. We end with formulating similar results for the parabolic case. After stating the results for the generic parabolic case we show that our model Hamiltonian H_{bif}

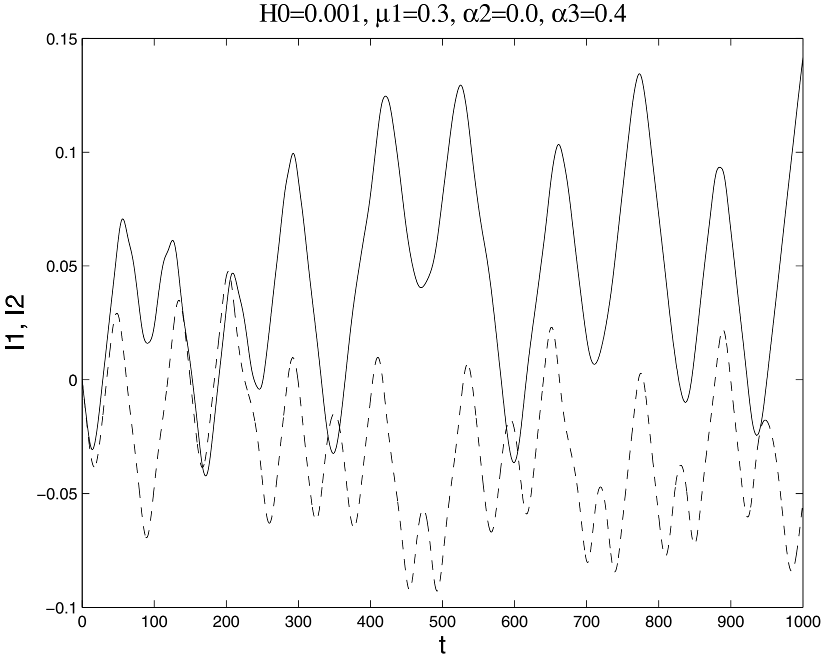


Figure 27. *Instability in the action variables near* 2*-PR: A perturbed orbit projected on the time-actions plane, with initial conditions as in Figure* 26*. Solid line:* I_1*. Dashed line:* I_2*.*

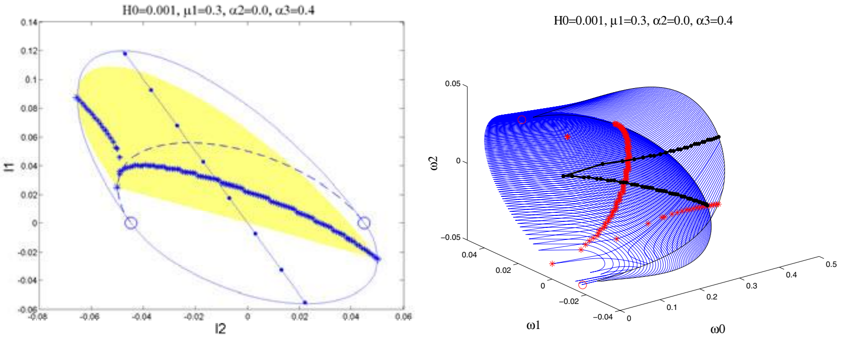


Figure 28. *An energy surface of* H_{bif} *with* $\alpha_2 = 0$*,* $H_0 = 1e - 3$*. Left: An EMBD in the* (I_2, I_1) *plane. Right: In the frequency space.*

is nongeneric in this context (because of its Z_2 symmetry) and formulate the corresponding results to a suitable class of Hamiltonians.

8.1. Folds in the EMBD and resonances.

Consider the EMBD near a singular family of $n-1$ lower dimensional tori, p_f. (Here we again take $s = 1$. General value of s will be considered elsewhere.) The unperturbed Hamiltonian, expressed in suitable local coordinates near p_f, is given by $H_0 = H_0(x, y, I)$, where $p_f = (x_f, y_f, I_f)$ satisfies $\nabla_{x,y} H_0(x, y, I)|_{p_f} = 0$. By the implicit function theorem (IFT), if the Hessian of H_0 with respect to x, y is nonsingular at p_f (namely, det $\frac{\partial^2 H(x,y,I)}{\partial x \partial y}|_{p_f} \neq 0$), we may express this manifold as a graph over the I variables: $p_f = (x_f(I), y_f(I), I)$. Then, apart from parabolic points (where det $\frac{\partial^2 H(x,y,I)}{\partial x \partial y}|_{p_f} = 0$), p_f is represented in the EMBD by the codimension one smooth manifold $p_f^h = \{h_f(I), I\} = \{H_0(x_f(I), y_f(I), I), I\}$.[15] It is now natural to define extremal points of the singularity surface in the EMBD.

Definition 5. *p_f^c is a simple k-extremal point of the singularity manifold p_f^h if p_f^c is nonparabolic and $h_f(I) = H_0(x_f(I), y_f(I), I)$ has a local extremal in k directions at p_f^c; i.e., there exist $i_1, \dots, i_k \in \{1, \dots, n-1\}$ such that*

$$\left.\frac{\partial h_f(I)}{\partial I_i}\right|_{p_f^c} = 0 \textit{ for } i = i_1, \dots, i_k. \tag{8.1}$$

Theorem 1. *Consider a family of singular nonparabolic $n-1$ dimensional tori $p_f(I) = (x_f(I), y_f(I), I)$, where (x, y, I) are suitable coordinates near $p_f(I)$. Then $p_f^c = p_f(I^c)$ is a simple k-extremal point of the corresponding singularity manifold in the EMBD iff p_f^c corresponds to a k-strongly resonant lower dimensional torus.*

Proof. Since p_f^c is not parabolic the representation $p_f^h = \{h_f(I) = H_0(x_f(I), y_f(I), I), I\}$ is nonsingular near p_f^c. Hence, p_f^c is a k-extremal point iff the surface $\{h_f(I) = H_0(x_f(I), y_f(I), I), I\}$ in the (h, I) space is extremal in k directions $I_{i_1}, \dots, I_{i_k}$ at p_f^c. This occurs, by definition, iff $\frac{\partial h_f(I)}{\partial I_i}$ vanishes in the corresponding k directions at p_f^c as expressed in (8.1). Since we use suitable coordinates, and since $\nabla_{x,y} H_0(x, y, I)|_{p_f} = 0$, it follows that for $i = i_1, \dots, i_k$

$$\left.\dot{\theta}_i\right|_{p_f^c} = \left.\frac{\partial H(x,y,I)}{\partial I_i}\right|_{p_f^c} = \left.\frac{dH(p_f(I))}{dI_i}\right|_{p_f^c} = \left.\frac{\partial h_f(I)}{\partial I_i}\right|_{p_f^c} = 0. \qquad \blacksquare \tag{8.2}$$

Theorem 1 relates the extremal point of the singularity surfaces in the EMBD and resonances. We have seen that the topology of the energy surfaces changes at folds of these singular surfaces. We note here the triviality that folds imply extremum points, and extremum points with first nonvanishing derivatives of even order imply folds.

8.2. Topological bifurcations. In the previous section we saw that H_{bif} has two values $h_c = h_{ell}^+, h_{hyp}^0$ which are simple 2-fold points of the elliptic and hyperbolic singularity surfaces (namely, these singularity surfaces have an even order extrema in two action directions), several families of curves on which a simple 1-fold occurs (corresponding to the intersection of the singularity surfaces $p_{ell}^\pm, p_{hyp}^0, p_{ell}^0$ with the corresponding resonances), and $h_c = h_p^0$ is a 1-parabolic fold point corresponding to the first appearance of parabolic tori. We observe that h_{ell}^+, h_{hyp}^0, and h_p^0 correspond to a topological change in the energy surfaces' structure, namely, the corresponding topology of the branched surfaces changes across these energy values, but that the families along which a simple 1-fold occurs do not correspond to such changes. We would like to formulate these observations. First, we need to define the branched surfaces in a precise way.

[15] With a slight abuse of notation we denoted it in previous sections by $p_f = (h_f(I), I)$ as well (see, for example, (7.2), (7.3), (7.4)).

Recall that h_c is a topological bifurcation point if the branched surfaces across h_c are not equivalent (definition 3). If the topology of A_h^S, the singularity manifold for a given energy surface, changes across h, then the branched surfaces across h are not equivalent. Using the Morse lemma we establish that for $s = 1$ the singularity manifold, A_h^S, changes its topology near folds of the singularity surfaces. Since folds of the singularity surfaces imply extrema and extrema imply resonances, the main result follows.

Definition 6. *p_f^c is an $n-k$ $(0<k<n)$ strongly resonant $n-1$ dimensional singular torus with nondegenerate frequency vector if in the suitable Arnold–Liouville–Nekhoroshev coordinates*

$$\frac{dH_0(p_f^c)}{dI_j} = 0, \quad \det\left(\frac{d^2H_0(p_f^c)}{dI_i dI_j}\right) \neq 0, \qquad j = 1,\dots,k. \tag{8.3}$$

The relation between (8.3) and resonances, as stated in the definition, follows from Theorem 1.

Theorem 2. *If p_f^c is a nonparabolic $n-1$ strongly resonant $n-1$ dimensional singular torus with a nondegenerate frequency vector, then $h_c = H_0(p_f^c)$ is a topological bifurcation point.*

Proof. The theorem essentially follows from the Morse lemma (see [26] or [41]); we include some details to enhance the intuition. Using the suitable coordinates near p_f, we may write

$$H_0(p_f - p_f^c) = H_0(p_f^c) + (I^f - I^c)^T A(I^f - I^c) + O(3), \tag{8.4}$$

where A is the Hessian at $p_f^c : A = \frac{d^2H_0(p_f^c)}{dI_i dI_j}$ (recall that $\nabla_{x,y}H_0(p_f - p_f^c) \equiv 0$). Hence, by linear orthonormal transformation $Uz = I$, we may write (8.4) as

$$H_0(p_f - p_f^c) - H_0(p_f^c) + \sum_{i=1}^{n-1-r} a_{i+r}z_{i+r}^2 = \sum_{i=1}^{r} a_i z_i^2 + O(3),$$

where $a_i > 0$ for all i by the nondegeneracy assumption. In fact, r is the Morse index of $h_f(I) = H_0(x_f(I), y_f(I), I)$ at p_f^c (the dimension of the subspace for which the Hessian A is positive definite). The Morse lemma, which applies to $h_f(I)$ by (8.3), states that by smooth local change of coordinates we can eliminate all higher order terms and set all the a_i's to unity. It follows immediately that intersection of the singularity surface $\{H_0(p_f), I_f\}$ with the plane $H_0(p_f) = h$ near p_f^c changes its topology across $h_c = H_0(p_f^c)$; if $r = n-1$ (resp., $r = 0$). Namely, if A is positive (negative) definite, then for $h < h_c$ $(h > h_c)$ there is no branch of p_f near p_f^c satisfying $H_0(p_f) = h$, whereas on the other side there is an $n-2$ dimensional ellipse satisfying this equation. If $0 < r < n-1$, the hyperboloids $p_f(h,\cdot)$ change their orientation at $h = h_c$; namely, they do not depend smoothly on h at h_c. Since the branched surfaces change across the surfaces $p_f(h,\cdot)$, the claim is proved. ■

8.3. Parabolic tori and topological bifurcations. Consider the surface of parabolic lower dimensional tori $p_{pf} = (x_{pf}, y_{pf}, I_{pf})$ so that

$$\nabla_{x,y}H_0(x,y,I)\big|_{p_{pf}} = \det \frac{\partial^2 H(x,y,I)}{\partial x \partial y}\bigg|_{p_{pf}} = 0. \tag{8.5}$$

These three equations define (generically) a codimension two surface in the EMBD, the singularity surface, p^h_{pf}. Along p^h_{pf} two (or more, in symmetric/degenerate cases) singularity surfaces representing families of nonparabolic $(n-1)$-tori, $p^h_{f_i}(I)$, $i = 1, \ldots, n$, meet. For example, in section 7, the four singularity surfaces $p^h_{f_{1,2}}(I) = p^{\pm}_{ell}$, $p^h_{f_{3,4}}(I) = p^0_{hyp,ell}$ meet at $p^h_{pf} = p^0_{par}$. The natural oscillations in the xy plane of the $n-1$ dimensional tori p_{pf} vanish, so these tori are strongly resonant in the ω_0-direction. We now address the natural question in view of Theorem 1: When do extremal points of this singularity surface (p^h_{pf}) in the EMBD correspond to additional strong resonances? Here, one should take careful limits when considering derivatives across the singular boundary of $p^h_{f_i}(I)$, namely, across p^h_{pf}. To formulate such conditions let us investigate more fully (8.5). Define the functions

$$f_1(x,y,I) = \frac{\partial H(x,y,I)}{\partial x}, \;\; f_2(x,y,I) = \frac{\partial H(x,y,I)}{\partial y}, \;\; f_3(x,y,I) = \det \frac{\partial^2 H(x,y,I)}{\partial x \partial y};$$

then (8.5) defines the surface $f_1(x,y,I) = f_2(x,y,I) = f_3(x,y,I) = 0$. Can this surface be represented as a graph over the $n-2$ actions $J^{n-2} = (I_1, \ldots, I_{j_p-1}, I_{j_p+1}, \ldots, I_{n-1})$ for some chosen index j_p? By the IFT, this may be done if $\frac{\partial(f_1,f_2,f_3)}{\partial(x,y,I_{j_p})}$ is nonsingular, and hence we have the following definition.

Definition 7. *p_{pf} is an $n-1$ dimensional parabolic torus which is nondegenerate in the I_{j_p} direction if p_{pf} satisfies* (8.5) *and*

$$\det \left. \frac{\partial\left(\frac{\partial H(x,y,I)}{\partial x}, \frac{\partial H(x,y,I)}{\partial y}, \det \frac{\partial^2 H(x,y,I)}{\partial x \partial y}\right)}{\partial(x,y,I_{j_p})} \right|_{p_{pf}} \neq 0. \tag{8.6}$$

If p^*_{pf} is an $n-1$ dimensional parabolic torus which is nondegenerate in the I_{j_p}-direction, then in its neighborhood there is an $n-2$ dimensional family of parabolic tori p_{pf} which may be expressed as a graph over the $n-2$ actions $J^{n-2} = (I_1, \ldots, I_{j_p-1}, I_{j_p+1}, \ldots, I_{n-1}) : p_{pf}(J^{n-2}) = (x_{pf}(J^{n-2}), y_{pf}(J^{n-2}), I_{j_p}(J^{n-2}), J^{n-2})$. It follows that the corresponding codimension two surface in the EMBD can be represented as a graph over the same actions as well:

$$\begin{aligned} p^h_{pf}(J^{n-2}) = \; & \{H_0(x_{pf}(J^{n-2}), y_{pf}(J^{n-2}), I_{j_p}(J^{n-2}), J^{n-2}), I_{j_p}(J^{n-2}), J^{n-2}\} \\ = \; & \{h_{pf}(J^{n-2}), I_{j_p}(J^{n-2}), J^{n-2}\}. \end{aligned}$$

Using (8.5), it follows that

$$\begin{aligned} \left.\frac{\partial h_{pf}(J^{n-2})}{\partial I_j}\right|_{p_{pf}} &= \left.\left(\frac{\partial H(x,y,I)}{\partial I_j} + \frac{\partial H(x,y,I)}{\partial I_{j_p}} \frac{\partial I_{j_p}(J^{n-2})}{\partial I_j}\right)\right|_{p_{pf}} \\ &= \left.\left(\dot{\theta}_j + \dot{\theta}_{j_p} \frac{\partial I_{j_p}(J^{n-2})}{\partial I_j}\right)\right|_{p_{pf}} \quad \text{for } j \neq j_p, \end{aligned} \tag{8.7}$$

whereas

$$\left.\dot{\theta}_{j_p}\right|_{p_{pf}} = \left.\frac{\partial H(x,y,I)}{\partial I_{j_p}}\right|_{p_{pf}} = \left.\frac{\partial h_{f_i}(I)}{\partial I_{j_p}}\right|_{p_{f_i} \to p_{pf}}, \tag{8.8}$$

and the independence of the last term on i follows from the smooth dependence of the Hamiltonian flow on I, even across the parabolic (in the xy-direction) point. The relation between extremal points of $h_{pf}(J^{n-2})$ (at which $\frac{\partial h_{pf}(I)}{\partial I_i}|_{p_{pf}}=0$) and resonances is now clear.

Theorem 3. *Consider a family of normally parabolic $n-1$ dimensional tori $p_{pf}=(x_{pf},y_{pf},I_{pf})$ which is nondegenerate in the I_{j_p}-direction at p^c_{pf}. Then, for $j\neq j_p$, an extremal point in the I_j-direction of the corresponding singularity manifold in the EMBD at p^c_{pf} corresponds to a strong resonance in this direction iff p^c_{pf} is strongly resonant in the I_{j_p}-direction or if I_{j_p} is extremal in I_j at p^c_{pf}. p^c_{pf} is strongly resonant in the I_{j_p}-direction iff the nonparabolic singularity surfaces emanating from p_{pf} are extremal in the I_{j_p}-direction in the limit $p_{f_i}\to p^c_{pf}$.*

An attempt to apply the above theorem to system (7.1) immediately fails—this system does not satisfy the nondegeneracy condition (8.6). Considering all systems with natural mechanical potential in the xy plane having a parabolic invariant $(n-1)$-torus at the origin,

$$H^{gen}_{bif}(x,y,I)=\frac{y^2}{2}-\frac{x^2}{2}f(I)+V(x,I), \tag{8.9}$$
$$f(0)=0,\qquad \frac{\partial V}{\partial x}\bigg|_{(0,0)}=0,\qquad \frac{\partial^2 V}{\partial x^2}\bigg|_{(0,0)}=0,$$

shows that the nondegeneracy condition (8.6) corresponds to

$$\frac{\partial^3 V}{\partial x^3}\frac{\partial^2 V}{\partial x\partial I_{j_p}}\bigg|_{(0,0)}\neq 0;$$

namely, the system is asymmetric with respect to reflections in x, and the location of the bifurcating invariant tori depends on I_{j_p}. Hence, any natural mechanical system with Z_2 symmetry does not satisfy (8.6).

We observe that another possibility (which is realized in our case of system (7.1)) of satisfying (8.5) along a simple $n-2$ dimensional surface is to require that the unperturbed system separates to a sum of two Hamiltonians, the first depending on $(x,y,\ I_{j_p})$ and the second depending on the actions $(I_1,\dots,I_{n-1})$. In this case equations (8.5) are independent of $J^{n-2}=(I_1,\dots,I_{j_p-1},I_{j_p+1},\dots,I_{n-1})$, and any solution of these equations is satisfied for all J^{n-2} values. This separability assumption is of course highly nongeneric from a mathematical point of view but is certainly of physical relevance (a similar approach appears in the theory of partial averaging).

Definition 8. *p_{pf} is an $n-1$ dimensional parabolic torus fully degenerate in the I_{j_p}-direction if p_{pf} does not satisfy (8.6) but does satisfy (8.5) and these equations are independent of I_j for all $j\neq j_p$.*

This condition is satisfied for any system of the form (8.9) if $f(I)=f(I_{j_p}),V(x,I)=V(x,I_{j_p})+g(I)$. In this case p_{pf} can be locally presented as the surface $(x_{pf},y_{pf},I_{j_p,pf},J^{n-2})$ with $x_{pf},y_{pf},I_{j_p,pf}$ independent of J^{n-2} so $\frac{\partial I_{j_p,pf}}{\partial I_j}=0$ for $j\neq j_p$. Indeed, for (7.1), we saw that $j_p=1$ and $p_{pf}=(x_{pf},y_{pf},I_{1,pf},I_2)=(0,0,0,I_2)$. Hence, near a parabolic torus, which is fully degenerate in the I_{j_p} direction, we can present the $n-2$ dimensional family of parabolic tori as $p^h_{pf}(J^{n-2})=\{H_0(x_{pf},y_{pf},I_{j_p},J^{n-2}),I_{j_p},J^{n-2}\}=\{h_{pf}(J^{n-2}),I_{j_p},J^{n-2}\}$. The relation between extrema of $h_{pf}(J^{n-2})$ and additional resonances follows immediately from (8.7) and (8.8), where we use $\frac{\partial I_{jp}}{\partial I_j} = 0$ to conclude the following.

Theorem 4. *Consider a family of normally parabolic $n-1$ dimensional tori $p_{pf} = (x_{pf}, y_{pf}, I_{pf})$, which is fully degenerate in the I_{j_p}-direction at p^c_{pf}. Then, for $j \neq j_p$, the extremal point in the I_j-direction of the corresponding singularity manifold in the EMBD at p^c_{pf} corresponds to a strong resonance in this direction. p^c_{pf} is strongly resonant in the I_{j_p}-direction iff the nonparabolic singularity surfaces emanating from p_{pf} are extremal in the I_{j_p}-direction in the limit $p_{f_i} \to p^c_{pf}$.*

As in section 8.2, by the Morse lemma, $n-2$ nondegenerate folds in the direction of the $n-2$ actions J^{n-2} of the codimension two surface $p^h_{pf}(J^{n-2})$ correspond to topological bifurcations. From (8.7) and (8.8) we conclude that such folds are not always associated with resonances, and we should distinguish between three cases.[16] For the fully degenerate case folds and resonances are simply related.

Theorem 5. *Consider an $n-1$ dimensional parabolic torus, p^c_{pf}. Assume p^c_{pf} is completely degenerate in the I_{j_p}-direction and that p^c_{pf} is $n-2$ strongly resonant with nondegenerate frequency vector in the $I_1, \dots, I_{j-1}, I_{j+1}, \dots, I_{n-1}$-directions; then $h_{pc} = H_0(p^c_{pf})$ is a topological bifurcation point.*

In section 7, the energy $h_{pc} = h^0_p$ is a topological bifurcation point which is well described by this theorem; at h^0_p parabolic tori first appear, and we have seen that a resonance in the I_2-direction occurs there.

In the generic case, a fold of $p^h_{pf}(J^{n-2})$ in the I_j-direction is associated with resonance if $\dot{\theta}_{j_p} = 0$ or if $\frac{\partial I_{j_p}(J^{n-2})}{\partial I_j}|_{p^c_{pf}} = 0$. Hence, topological bifurcations occurring at an $n-2$ fold of $p^h_{pf}(J^{n-2})$ are associated with a resonance only if additional conditions are satisfied. To satisfy these additional conditions in a persistent way the system must have additional parameters or symmetries.

Theorem 6. *Consider an $n-1$ dimensional parabolic torus, p^c_{pf}. Assume that p^c_{pf} is nondegenerate in the I_{j_p}-direction and that the Hamiltonian at p^c_{pf} is locally separable, namely, $\frac{\partial I_{j_p}(J^{n-2})}{\partial I_j}|_{p^c_{pf}} = 0$ for all $j \neq j_p$. Then, if p^c_{pf} is $n-2$ strongly resonant with nondegenerate frequency vector in the $I_1, \dots, I_{j-1}, I_{j+1}, \dots, I_{n-1}$-directions, then $h_{pc} = H_0(p^c_{pf})$ is a topological bifurcation point. Without imposed symmetries, such a phenomenon is of codimension $n-2$.*

Theorem 7. *If p^c_{pf} is an $n-1$ dimensional parabolic torus of fixed points which is nondegenerate in the I_{j_p}-direction and has nondegenerate frequency vector in the $I_1, \dots, I_{j-1}, I_{j+1}, \dots, I_{n-1}$-directions then $h_{pc} = H_0(p^c_{pf})$ is a topological bifurcation point. Without imposed symmetries, such a phenomenon is of codimension one.*

9. Discussion.

We have shown that when the generalized action-angle coordinates can be extended globally (as in our prototype models of normally stable, unstable, and bifurcating

[16] Note that there are two different (independent) types of nondegeneracies. One corresponds to the standard assumption regarding changes in the frequency vector (Definition 6) and is needed for applying the Morse lemma. The other corresponds to the nondegenerate (degenerate) dependence of the parabolic tori on a specific action (Definitions 7,8).

tori) the combination of the EMBD and the branched surfaces supply global qualitative description of the *near-integrable* dynamics; on these diagrams the topological changes in the energy surfaces and the appearance of lower dimensional resonances are apparent, and thus various mechanisms for instabilities (such as homoclinic orbits, hyperbolic resonances, and PR) may be clearly identified. In particular, we proved that topological bifurcations of the energy surfaces correspond to folds of singularity surfaces in these diagrams and hence to resonances. In other works [37, 38] we have demonstrated that the curvature of these singularity surfaces at the folds plays a crucial role in the extent of the instabilities in the perturbed system. Again, such effects are easily identified in these diagrams.

Many issues remain for future studies:

- We have seen (sections 6.3 and 7.3) that there is a long list of instabilities associated with the near-integrable motion near families of lower dimensional tori which is not well understood yet.
- For 2 DOF systems, the description of the energy surfaces as graphs gives a useful insight regarding the evolution of the instabilities in the action variables (or, more generally, in the adiabatic variables of the system) under small conservative perturbations or conservative noise [20]. These ideas were generalized to n DOF systems with strong conservative noise which destroys all integrals of motion and small nonconservative noise which leads to diffusion between different energy surfaces [19]. In view of our work, one is lead naturally to investigation of motion in integrable (or near-integrable) systems with small conservative noise by studying *random motion along branched surfaces*.
- The behavior of systems for which the local generalized action-angle coordinates cannot be globally extended is yet to be studied. In particular, one would like to extend the presentation here so it will be applicable to the work of Fomenko and coworkers in which the topology of complicated systems, like the rigid body, is fully analyzed [17, 18]. On one hand, one may use general constants of motion plots in a similar fashion to what we have proposed for the EMBD, yet the relation between folds and resonances will be lost. On the other hand, even for such plots, finding the branched surfaces topology from the Fomenko graphs is challenging.
- The restriction to systems with compact level sets excludes important examples such as the Kepler problem. Delicate issues related to the possible appearance of noncompact critical level sets and singularities of the potential need to be addressed (see [47]).
- Notice that the generalized action-angle local representation naturally leads to investigation of the Hamiltonian function evaluated along the singularities as a function of the $n-s$ actions. Hence, as noted in [33], singularity theory may be used to classify all persistent bifurcations in the s DOF subsystem. Here, we further observe that resonances are also associated with singularities of this function. Full classification, as had been achieved for some of the bifurcation scenarios, is yet to be developed.
- Finally, the effect of $n-s$ dimensional tori with various stabilities in the $2s$ dimensional normal space for $s > 1$ (as in [33]) on the EMBD structure and the branched surfaces structure is yet to be understood.

Acknowledgments. It is a pleasure to thank Amadeu Delshams, Heinz Hanßmann, Rafael de la Llave, and Terre Seara for fruitful discussions.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings Publishing, Reading, MA, 1978.

[2] V. I. Arnol'd, *Dynamical Systems* III, 2nd ed., Encyclopaedia Math. Sci. 3, Springer-Verlag, Berlin, 1993.

[3] S. V. Bolotin and D. V. Treschev, *Remarks on the definition of hyperbolic tori of Hamiltonian systems*, Regul. Chaotic Dyn., 5 (2000), pp. 401–412.

[4] A. V. Bolsinov and A. T. Fomenko, *Trajectory classification of simple integrable Hamiltonian systems on three-dimensional surfaces of constant energy*, Dokl. Akad. Nauk, 332 (1993), pp. 553–555 (in Russian). English translation in Russian Acad. Sci. Dokl. Math., 48, (1994), pp. 365–369.

[5] H. W. Broer, G. B. Huitema, and M. B. Sevryuk, *Quasi-Periodic Tori in Families of Dynamical Systems: Order Amidst Chaos*, Lecture Notes in Math. 1645, Springer-Verlag, New York, 1996.

[6] R. Cushman, *The momentum mapping of the harmonic oscillator*, in Symposia Mathematica, Vol. 14, Academic Press, London, 1974, pp. 323–342.

[7] R. H. Cushman and L. M. Bates, *Global Aspects of Classical Integrable Systems*, Birkhäuser Verlag, Basel, 1997.

[8] A. Delshams, R. de la Llave, and T. M. Seara, *A geometric approach to the existence of orbits with unbounded energy in generic periodic perturbations by a potential of generic geodesic flows of* $\mathbb{T}^2$, Comm. Math. Phys., 209 (2000), pp. 353–392.

[9] A. Delshams, R. de la Llave, and T. M. Seara, *A geometric mechanism for diffusion in Hamiltonian systems overcoming the large gap problem: Announcement of results*, Electron. Res. Announc. Amer. Math. Soc., 9 (2003), pp. 125–134.

[10] H. R. Dullin, M. Juhnke, and P. H. Richter, *Action integrals and energy surfaces of the Kovalevskaya top*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1535–1562.

[11] H. R. Dullin, P. H. Richter, and A. P. Veselov, *Action variables of the Kovalevskaya top*, Regul. Chaotic Dyn., 3 (1998), pp. 18–26.

[12] L. H. Eliasson, *Perturbations of stable invariant tori for Hamiltonian systems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 15 (1988), pp. 115–147.

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

[14] N. Fenichel, *Asymptotic stability with rate conditions*, Indiana Univ. Math. J., 23 (1974), pp. 1109–1137.

[15] N. Fenichel, *Asymptotic stability with rate conditions,* II, Indiana Univ. Math. J., 26 (1977), pp. 81–93.

[16] A. T. Fomenko, *Topological classification of all integrable Hamiltonian differential equations of general type with two degrees of freedom*, in The Geometry of Hamiltonian Systems, Math. Sci. Res. Inst. Publ. 22, Springer-Verlag, New York, 1991, pp. 131–339.

[17] A. T. Fomenko, Ed., *Topological classification of integrable systems*, Advances in Soviet Mathematics 6, AMS, Providence, RI, 1991.

[18] A. T. Fomenko, Ed., *Symplectic topology of integrable dynamical systems. Rough topological classification of classical cases of integrability in the dynamics of a heavy rigid body*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 235 (1996), pp. 104–183, 305. Translation in J. Math. Sci. (New York), 94 (1999), pp. 1512–1557.

[19] M. Freidlin and M. Weber, *On random perturbations of Hamiltonian systems with many degrees of freedom*, Stochastic Process. Appl., 94 (2001), pp. 199–239.

[20] M. I. Freidlin and A. D. Wentzell, *Random perturbations of Hamiltonian systems*, Mem. Amer. Math. Soc., 109 (1994).

[21] S. M. Graff, *On the conservation of hyperbolic invariant tori for Hamiltonian systems*, J. Differential Equations, 15 (1974), pp. 1–69.

[22] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[23] G. Haller, *Chaos Near Resonance*, Appl. Math. Sci. 138, Springer-Verlag, New York, 1999.

[24] H. Hanßmann, *The quasi-periodic center-saddle bifurcation*, J. Differential Equations, 142 (1997), pp. 305–370.

[25] H. Hanßmann, *A survey on bifurcations of invariant tori*, in New Advances in Celestial Mechanics and Hamiltonian Systems, J. Delgado et al., eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004, pp. 109–121.

[26] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Encyclopedia Math. Appl. 54, Cambridge University Press, Cambridge, UK, 1995.

[27] J. Laskar, *Frequency analysis for multi-dimensional systems. global dynamics and diffusion*, Phys. D, 67 (1993), pp. 257–281.

[28] J. Laskar, *Global dynamics and diffusion*, Phys. D, 67 (1993), pp. 257–281.

[29] L. Lerman, *Isoenergetical structure of integrable Hamiltonian systems in an extended neighborhood of a simple singular point: Three degrees of freedom*, Amer. Math. Soc. Transl. Ser. 2, 200 (2000), pp. 219–242.

[30] L. Lerman and Y. L. Umanskii, *Classification of four dimensional integrable Hamiltonian systems and Poisson actions of $\mathbb{R}^2$ in extended neighborhood of simple singular points,* I, Sb. Math., 77 (1994), pp. 511–542.

[31] L. Lerman and Y. L. Umanskii, *Classification of four dimensional integrable Hamiltonian systems and Poisson actions of $\mathbb{R}^2$ in extended neighborhood of simple singular points,* II, Sb. Math., 78 (1994), pp. 479–506.

[32] L. M. Lerman and Y. L. Umanskii, *Classification of four-dimensional integrable Hamiltonian systems and Poisson actions of $\mathbb{R}^2$ in extended neighborhoods of simple singular points.* III. *Realizations*, Mat. Sb., 186 (1995), pp. 89–102 (in Russian). English translation in Sb. Math., 186 (1995), pp. 1477–1491.

[33] L. M. Lerman and Y. L. Umanskii, *Four-dimensional integrable Hamiltonian systems with simple singular points (topological aspects)*, Transl. Math. Monogr. 176, AMS, Providence, RI, 1998. Translated from the Russian manuscript by A. Kononenko and A. Semenovich.

[34] A. Lichtenberg and M. Lieberman, *Regular and Stochastic Motion*, Appl. Math. Sci. 38, Springer-Verlag, New York, 1983.

[35] A. Litvak-Hinenzon, *Parabolic Resonances in Hamiltonian Systems*, Ph.D. thesis, The Weizmann Institute of Science, Rehovot, Israel, 2001.

[36] A. Litvak-Hinenzon and V. Rom-Kedar, *Parabolic resonances in near integrable Hamiltonian systems*, in Stochaos: Stochastic and Chaotic Dynamics in the Lakes, D. S. Broomhead, E. A. Luchinskaya, P. V. E. McClintock, and T. Mullin, eds., American Institute of Physics, Melville, NY, 2000, pp. 358–368.

[37] A. Litvak-Hinenzon and V. Rom-Kedar, *Parabolic resonances in* 3 *degree of freedom near-integrable Hamiltonian systems*, Phys. D, 164 (2002), pp. 213–250.

[38] A. Litvak-Hinenzon and V. Rom-Kedar, *Resonant tori and instabilities in Hamiltonian systems*, Nonlinearity, 15 (2002), pp. 1149–1177.

[39] A. Litvak-Hinenzon, *The Mechanism of Parabolic Resonance*, in Equadiff 2003, Hasselt, Belgium, J. Mawhin and S. V. Lunel, eds., World Scientific, Singapore, to appear.

[40] K. R. Meyer and R. G. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Appl. Math. Sci. 90, Springer-Verlag, New York, 1991.

[41] J. Milnor, *Morse Theory. Based on Lecture Notes by M. Spivak and R. Wells*, Ann. of Math. Stud. 51, Princeton University Press, Princeton, NJ, 1963.

[42] N. N. Nehorošev, *Action-angle variables, and their generalizations*, Trans. Moscow Math. Soc., 26 (1972), pp. 181–198.

[43] A. A. Oshemkov, *Description of isoenergetic surfaces of integrable Hamiltonian systems with two degrees of freedom*, Trudy Sem. Vektor. Tenzor. Anal., 23 (1988), pp. 122–132 (in Russian).

[44] J. Pöschel, *On elliptic lower dimensional tori in Hamiltonian systems*, Math. Z., 202 (1989), pp. 559–608.

[45] V. Rom-Kedar, *Parabolic resonances and instabilities*, Chaos, 7 (1997), pp. 148–158.

[46] V. Rom-Kedar, Y. Dvorkin, and N. Paldor, *Chaotic Hamiltonian dynamics of particle's horizontal motion in the atmosphere*, Phys. D, 106 (1997), pp. 389–431.

[47] S. Smale, *Topology and mechanics.* I, Invent. Math., 10 (1970), pp. 305–331.

[48] J. L. Tennyson, M. A. Lieberman, and A. J. Lichtenberg, *Diffusion in near-integrable Hamiltonian systems with three degrees of freedom*, in Nonlinear Dynamics and the Beam-Beam Interaction AIP Conf. Proc. 57, Amer. Inst. Phys., New York, 1980, pp. 272–301.

[49] D. Treschev, *Multidimensional symplectic separatrix maps*, J. Nonlinear Sci., 12 (2002), pp. 27–58.

[50] D. Treschev, *Trajectories in a neighbourhood of asymptotic surfaces of a priori unstable Hamiltonian systems*, Nonlinearity, 15 (2002), pp. 2033–2052.

[51] H. Waalkens and H. R. Dullin, *Quantum monodromy in prolate ellipsoidal billiards*, Ann. Physics, 295 (2002), pp. 81–112.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 574–600

Evans Functions for Integral Neural Field Equations with Heaviside Firing Rate Function*

S. Coombes[†] and M. R. Owen[†]

Abstract. In this paper we show how to construct the Evans function for traveling wave solutions of integral neural field equations when the firing rate function is a Heaviside. This allows a discussion of wave stability and bifurcation as a function of system parameters, including the speed and strength of synaptic coupling and the speed of axonal signals. The theory is illustrated with the construction and stability analysis of front solutions to a scalar neural field model, and a limiting case is shown to recover recent results of Zhang [*Differential Integral Equations*, 16 (2003), pp. 513–536]. Traveling fronts and pulses are considered in more general models possessing either a linear or piecewise constant recovery variable. We establish the stability of coexisting traveling fronts beyond a front bifurcation and consider parameter regimes that support two stable traveling fronts of different speeds. Such fronts may be connected, and depending on their relative speed the resulting region of activity can widen or contract. The conditions for the contracting case to lead to a pulse solution are established. The stability of pulses is obtained for a variety of examples, in each case confirming a previously conjectured stability result. Finally, we show how this theory may be used to describe the dynamic instability of a standing pulse that arises in a model with slow recovery. Numerical simulations show that such an instability can lead to the shedding of a pair of traveling pulses.

Key words. traveling waves, neural networks, integral equations, Evans functions

AMS subject classification. 92C20

DOI. 10.1137/040605953

1. Introduction.

Traveling waves in neurobiology are receiving increased attention from experimentalists, in part due to their ability to visualize them with multielectrode recordings and imaging methods. In particular, it is possible to electrically stimulate slices of pharmacologically treated tissue taken from the cortex [19], hippocampus [31], and thalamus [28]. For cortical circuits such in vitro experiments have shown that, when stimulated appropriately, waves of excitation may occur [12, 40]. Such waves are a consequence of synaptic interactions and the intrinsic behavior of local neuronal circuitry. The propagation speed of these waves is of the order $\mathrm{cm\,s^{-1}}$, an order of magnitude slower than that of action potential propagation along axons. The class of computational models that are believed to support synaptic waves differ radically from classic models of waves in excitable systems. Most importantly, synaptic interactions are nonlocal (in space) and involve communication (space-dependent) delays (arising from the finite propagation velocity of an action potential) and distributed delays (arising from neurotransmitter release and dendritic processing). In many continuum models

[*] Received by the editors March 30, 2004; accepted for publication (in revised form) by D. Terman July 9, 2004; published electronically November 9, 2004.

http://www.siam.org/journals/siads/3-4/60595.html

[†] Department of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK (stephen.coombes@nottingham.ac.uk, markus.owen@nottingham.ac.uk). S. Coombes would like to acknowledge ongoing support from the EPSRC through the award of an Advanced Research Fellowship, grant GR/R76219.

for the propagation of electrical activity in neural tissue it is assumed that the synaptic input current is a function of the presynaptic firing rate function [39]. These infinite dimensional dynamical systems typically take the form [15]

$$\frac{1}{\alpha}\frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{-\infty}^{\infty} \mathrm{d}y w(y) f \circ u(x-y,t) - ga(x,t), \tag{1.1}$$

$$\frac{1}{\epsilon}\frac{\partial a(x,t)}{\partial t} = -a(x,t) + u(x,t). \tag{1.2}$$

Here, $u(x,t)$ is interpreted as a neural field representing the local activity of a population of neurons at position $x \in \mathbb{R}$, $f \circ u$ denotes the firing rate function, and $w(y)$ denotes the strength of connections between neurons separated by a distance y (assuming a spatially homogeneous system). The constant α is the synaptic rate constant, while the neural field $a(x,t)$ represents a local feedback signal, with strength $g > 0$, that modulates synaptic currents. Numerical simulations, with sigmoidal f, show that such systems support unattenuated traveling waves as a result of localized input. In the absence of local feedback dynamics (i.e., $g = 0$), Ermentrout and McLeod [16] have established that there exists a unique monotone traveling front solution for sigmoidal firing rate functions and positive spatially decaying weight functions. Indeed, there are now a number of results about existence, uniqueness, and asymptotic stability for integral differential equations, such as can be found in [2, 10, 11, 9]. Recently Pinto and Ermentrout [34, 35] have constructed traveling pulse solutions using a singular perturbation argument for a general class of continuous firing rate functions. Moreover, for the special choice of a Heaviside firing rate function they have made extensive use of techniques pioneered by Amari [2] for the explicit construction of bumps and waves. For the case of standing pulses (and a more general form of local feedback than (1.2)) they have also given a rigorous analysis of the linearized equations of motion, allowing a discussion of bump stability. In this paper we extend this work to cover the stability of traveling (as well as standing) waves, using an Evans function approach. Moreover, we shall consider a general class of neural field theories that contains standard models, such as (1.1) and (1.2), as limiting cases.

The Evans function is a powerful tool for the stability analysis of nonlinear waves on unbounded domains. It has previously been used within the context of PDEs but has also been recently formulated for neural field theories [41] and more general nonlocal problems [25]. The main use of the Evans function is in locating the point spectrum (isolated eigenvalues) of some relevant linearized operator. While its computation is typically only possible in perturbative situations, Zhang [41] has found an explicit formula for the Evans function for traveling waves of (1.1) and (1.2) with a Heaviside firing rate function. His approach makes explicit use of the inhomogeneous ordinary differential equation structure of the linearized equations of motion around a traveling wave. These are solved via the method of variation of parameters, with the source term arising from the nonlocal nature of the model. In this way he first constructs an intermediate Evans function for the homogeneous problem before using this to determine the full Evans function in an appropriate right half plane. From here he is able to establish that the fast traveling pulse solution of the singularly perturbed system (1.1) and (1.2) with $\epsilon \ll \alpha$ is exponentially stable. By working with integral rather than integro-differential models, we shall develop three important extensions of this approach, side-stepping the need to construct an intermediate Evans function. These extensions cover

(i) the study of exponential wave stability in an integral framework rather than an integro-differential framework, which contains models like (1.1) and (1.2) as special cases, (ii) avoiding the need to resort to the study of some singularly perturbed system, and (iii) including the effects of space-dependent delays arising from axonal communication.

In section 2 we introduce the form of integral neural field model that we are concerned with in this paper. To solve the stability problem of the traveling wave solution, we rewrite the integral equations in moving coordinates and linearize about the traveling wave. Special solutions of this linearized system give rise to an eigenvalue problem and a linear operator $\mathcal{L}$. To establish the exponential stability of traveling wave solutions it suffices to investigate the spectrum of this operator. Since we are concerned with systems where the real part of the continuous spectrum has a uniformly negative upper bound, it is vitally important to determine the location of the isolated spectrum for wave stability. For the case of a Heaviside firing rate function we show in section 3 how this spectrum may be determined by the zeros of a complex analytic function, which we identify as the Evans function. For illustrative purposes the focus of this section is on scalar integral models with traveling front solutions. In the next two sections we consider models with linear and nonlinear recovery variables, respectively, that can also support traveling pulses. Throughout sections 3, 4, and 5 a number of examples are presented to illustrate the application of this theory. Moreover, we are able to establish a number of previously conjectured stability results. Finally, in section 6 we discuss natural extensions of this work.

2. Traveling waves.

In this section we introduce a more general integral form of (1.1) and consider the analysis of traveling wave solutions. For clarity we reserve discussion of feedback until later sections and take $g = 0$. Apart from a spatial integral mixing of the network connectivity function with space-dependent delays, arising from noninstantaneous axonal communication, integral models can naturally incorporate a temporal integration over some appropriately identified distributed delay kernel. These distributed delay kernels are biologically motivated and represent the response of (slow) biological synapses to spiking inputs. For a general firing rate function of synaptic current we consider the scalar integral equation

$$u(x,t) = \int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_0^{\infty} \mathrm{d}s \eta(s) f \circ u(x-y, t-s-|y|/v). \tag{2.1}$$

Here v represents the velocity of action potential propagation [39, 23], while $\eta(t)$ ($\eta(t) = 0$ for $t < 0$) models the effects of synaptic processing. With the choice $\eta(t) = \alpha \mathrm{e}^{-\alpha t}$ we recover the model given by (1.1). Some discussion of traveling wave solutions to (2.1) has previously been given in [13], where for some specific choices of $w(x)$ the equivalence to a certain PDE was exploited. Throughout this paper we shall avoid the use of such techniques and always work with integral equation models directly. However, again for simplicity, we shall consider the restriction $w(x) = w(|x|)$.

Following the standard approach for constructing traveling wave solutions to PDEs, such as reviewed by Sandstede [36], we introduce the coordinate $\xi = x - ct$ and seek functions $U(\xi, t) = u(x - ct, t)$ that satisfy (2.1). In the (ξ, t) coordinates, the integral equation (2.1) reads

$$U(\xi,t) = \int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_0^{\infty} \mathrm{d}s \eta(s) f \circ U(\xi - y + cs + c|y|/v, t - s - |y|/v). \tag{2.2}$$

The traveling wave is a stationary solution $U(\xi,t) = q(\xi)$ (independent of t) that satisfies

$$q(\xi) = \int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_0^{\infty} \mathrm{d}s \eta(s) f \circ q(\xi - y + cs + c|y|/v). \tag{2.3}$$

The linearization of (2.2) about the steady state $q(\xi)$ is obtained by writing $U(\xi,t) = q(\xi) + u(\xi,t)$, and Taylor expanding, to give

$$\begin{aligned} u(\xi,t) = \int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_0^{\infty} \mathrm{d}s \eta(s) f'(q(\xi - y + cs + c|y|/v)) \\ \times u(\xi - y + cs + c|y|/v, t - s - |y|/v). \end{aligned} \tag{2.4}$$

Of particular importance are bounded smooth solutions defined on $\mathbb{R}$ for each fixed t. Thus one looks for solutions of the form $u(\xi,t) = u(\xi)\mathrm{e}^{\lambda t}$. This leads to the eigenvalue equation $u = \mathcal{L}u$:

$$\begin{aligned} u(\xi) = \int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_{\xi - y + c|y|/v}^{\infty} \frac{\mathrm{d}s}{c} \eta(-\xi/c + y/c - |y|/v + s/c) \\ \times \mathrm{e}^{-\lambda(-\xi/c + y/c + s/c)} f'(q(s))u(s). \end{aligned} \tag{2.5}$$

Let $\sigma(\mathcal{L})$ be the spectrum of $\mathcal{L}$. We shall say that a traveling wave is linearly stable if

$$\max\{\mathrm{Re}(\lambda) : \lambda \in \sigma(\mathcal{L}),\ \lambda \neq 0\} \leq -K \tag{2.6}$$

for some $K > 0$, and $\lambda = 0$ is a simple eigenvalue of $\mathcal{L}$. Furthermore, we shall take it that linear stability implies nonlinear stability. Although this is known to be true in a broad range of cases, including for asymptotically constant traveling waves in both reaction-diffusion equations and viscous conservation laws (see, for example, [21]), it is an open challenge to establish this for the integral equations that we consider in this paper.

In general the normal spectrum of the operator obtained by linearizing a system about its traveling wave solution may be associated with the zeros of a complex analytic function, the so-called Evans function. This was originally formulated by Evans [17] in the context of a stability theorem about excitable nerve axon equations of Hodgkin–Huxley type. Jones subsequently employed this function with some geometric techniques to establish the stability of fast traveling pulses in the FitzHugh–Nagumo model [24]. Following from this, Alexander, Gardner, and Jones formulated a more general method to define the Evans function for semilinear parabolic systems [1]. Indeed this approach has proved quite versatile and has now been used to study the stability of traveling waves in a number of PDE models, such as discussed in [32, 4, 8, 36]. The extension to integral models is far more recent [41, 25]. However, it is fair to say that although there are many nonlinear evolution equations which support traveling wave solutions, there are very few which possess explicit Evans functions

[33], except perhaps when the underlying PDE is integrable [26, 27, 30]. It is therefore all the more interesting that work by Zhang [41] on integral neural field models has shown that such explicit formulas are possible for the choice of a Heaviside firing rate function. This is one reason for us to pursue the choice of a Heaviside firing rate function. Another being that the qualitative features of traveling wave solutions for smooth sigmoidal firing rate functions, often used to describe biological firing rates, have previously been found to carry over to the case with a nonsmooth Heaviside firing rate function [16, 34, 13]. Throughout the rest of this paper we shall therefore focus on the choice $f(u) = \Theta(u-h)$ for some threshold h. In this case the traveling wave is given by

$$q(\xi) = \int_0^\infty \eta(s)\psi(\xi+cs)\mathrm{d}s, \tag{2.7}$$

where

$$\psi(\xi) = \int_{-\infty}^\infty w(y)\Theta(q(\xi-y+c|y|/v)-h)\mathrm{d}y, \tag{2.8}$$

with $f'(q) = \delta(q-h)$. Note that this has a legitimate interpretation as $f'(q)$ only ever appears within an integral, such as (2.5). In the next section we will describe how to construct the Evans function for traveling wave solutions given by (2.7) and (2.8).

3. Fronts in a scalar model.

In this section we introduce the techniques for constructing the Evans function with the example of traveling front solutions to (2.1). Previous work on the properties of such traveling fronts, i.e., speed as a function of system parameters, but not stability, can be found in [23, 34, 13]. Note also that a formal link between traveling front solutions in neural field theories and traveling spikes in integrate-and-fire networks can be found in [14]. We look for traveling front solutions such that $q(\xi) > h$ for $\xi < 0$ and $q(\xi) < h$ for $\xi > 0$. It is then a simple matter to show that

$$\psi(\xi) = \begin{cases} \int_{\xi/(1-c/v)}^\infty w(y)\mathrm{d}y, & \xi \geq 0, \\ \int_{\xi/(1+c/v)}^\infty w(y)\mathrm{d}y, & \xi < 0. \end{cases} \tag{3.1}$$

The choice of origin, $q(0) = h$, gives an implicit equation for the speed of the wave as a function of system parameters.

The construction of the Evans function begins with an evaluation of (2.5). Under the change of variables $z = q(s)$ this equation may be written

$$\begin{aligned} u(\xi) = \int_{-\infty}^\infty \mathrm{d}y w(y) \int_{q(\xi-y+c|y|/v)}^{q(\infty)} \frac{\mathrm{d}z}{c} \eta(q^{-1}(z)/c - \xi/c + y/c - |y|/v) \\ \mathrm{e}^{-\lambda(q^{-1}(z)/c-\xi/c+y/c)} \times \frac{\delta(z-h)}{|q'(q^{-1}(z))|} u(q^{-1}(z)). \end{aligned} \tag{3.2}$$

For the traveling front of choice we note that when $z = h$, $q^{-1}(h) = 0$ and (3.2) reduces to

$$u(\xi) = \frac{u(0)}{c|q'(0)|} \int_{-\infty}^\infty \mathrm{d}y w(y) \eta(-\xi/c + y/c - |y|/v) \mathrm{e}^{-\lambda(y-\xi)/c}. \tag{3.3}$$

From this equation we may generate a self-consistent equation for the value of the perturbation at $\xi = 0$, simply by setting $\xi = 0$ on the left-hand side of (3.3). This self-consistent condition reads

$$u(0) = \frac{u(0)}{c|q'(0)|} \int_{-\infty}^{\infty} \mathrm{d}y w(y) \eta(y/c - |y|/v) \mathrm{e}^{-\lambda y/c}. \tag{3.4}$$

Importantly there are only nontrivial solutions if $\mathcal{E}(\lambda) = 0$, where

$$\mathcal{E}(\lambda) = 1 - \frac{1}{c|q'(0)|} \int_{-\infty}^{\infty} \mathrm{d}y w(y) \eta(y/c - |y|/v) \mathrm{e}^{-\lambda y/c}. \tag{3.5}$$

From causality $\eta(t) = 0$ for $t \leq 0$ and physically $c < v$, so

$$\mathcal{E}(\lambda) = 1 - \frac{1}{c|q'(0)|} \int_{0}^{\infty} \mathrm{d}y w(y) \eta(y/c - y/v) \mathrm{e}^{-\lambda y/c}. \tag{3.6}$$

We identify (3.6) with the Evans function for the traveling front solution of (2.1). The Evans function is real-valued if λ is real. Furthermore, (i) the complex number λ is an eigenvalue of the operator $\mathcal{L}$ if and only if $\mathcal{E}(\lambda) = 0$, and (ii) the algebraic multiplicity of an eigenvalue is equal to the order of the zero of the Evans function. To establish the last two results requires some further work, which we briefly discuss.

Consider for the moment the case that $\eta(t)$ is an exponential. We do this without significant loss of generality since the discussion below generalizes to the case that $\eta(t)$ is the Green's function of a linear differential operator. For the choice $\eta(t) = \alpha \mathrm{e}^{-\alpha t}$, we may rewrite (3.3) in the form

$$u(\xi) = u(0) \left[1 - \frac{\alpha}{c|q'(0)|} \int_{0}^{c_\pm \xi/c} \mathrm{d}y w(y) \mathrm{e}^{-\alpha y/c_\pm} \mathrm{e}^{-\lambda y/c} \right] \mathrm{e}^{(\lambda+\alpha)\xi/c}, \tag{3.7}$$

where $c_\pm^{-1} = c^{-1} \mp v^{-1}$ and we take c_- when $\xi < 0$ and c_+ when $\xi > 0$. Taking the limit as $\xi \to \infty$ gives

$$\lim_{\xi \to \infty} u(\xi) = u(0) \mathcal{E}(\lambda) \lim_{\xi \to \infty} \mathrm{e}^{(\lambda+\alpha)\xi/c}. \tag{3.8}$$

Assuming that $\mathrm{Re}\, \lambda > -\alpha$ (which we shall see shortly is to the right of the essential spectrum), then $u(\xi)$ will be unbounded as $\xi \to \infty$ unless $\mathcal{E}(\lambda) = 0$. This is precisely our defining equation for an eigenvalue. Hence, $\mathcal{E}(\lambda) = 0$ if and only if λ is an eigenvalue. To prove (ii) is more involved, and as such we merely present the key component of a more rigorous proof. Differentiating both sides of (3.7) with respect to λ and taking the large ξ limit give

(3.9)
$$-\frac{1}{c} u(0) \mathcal{E}(\lambda) \xi + \mathrm{e}^{-(\lambda+\alpha)/c} \frac{\mathrm{d}u(\xi)}{\mathrm{d}\lambda} = u(0) \mathcal{E}'(\lambda) + \frac{\mathrm{d}u(0)}{\mathrm{d}\lambda} \left[1 - \frac{\alpha}{c|q'(0)|} \int_{0}^{c_\pm \xi/c} \mathrm{d}y w(y) \mathrm{e}^{-\alpha y/c_\pm} \mathrm{e}^{-\lambda y/c} \right].$$

For the case that the order of the zero of the Evans function is two, defined by the conditions $\mathcal{E}(\lambda)=0$ and $\mathcal{E}'(\lambda)=0$ we have (in the large ξ limit) simply that

$$\frac{\mathrm{d}u(\xi)}{\mathrm{d}\lambda}=\frac{\mathrm{d}u(0)}{\mathrm{d}\lambda}\left[1-\frac{\alpha}{c|q'(0)|}\int_0^{c_\pm\xi/c}\mathrm{d}y w(y)\mathrm{e}^{-\alpha y/c_\pm}\mathrm{e}^{-\lambda y/c}\right]\mathrm{e}^{(\lambda+\alpha)\xi/c}, \tag{3.10}$$

which is the same as (3.7) under the replacement $u\rightarrow \mathrm{d}u/\mathrm{d}\lambda$. This shows that for the same eigenvalue there are two solutions (at least in the large ξ limit). Hence, a repeated root of the Evans function leads to an algebraic doubling of the eigenvalue. In a similar fashion mth order roots of the Evans function can be shown to lead to m distinct solutions with a common eigenvalue. These solutions are linear combinations of $\mathrm{d}^l u/\mathrm{d}\lambda^l$, $l=0,\dots,m$. Although we do not attempt to do so here, a rigorous proof of (ii) may be developed based around this idea.

From (2.7) and (2.8) it is straightforward to calculate $q'(\xi)$ in the form

$$q'(\xi)=\int_{-\infty}^{\infty}\mathrm{d}y w(y)\int_0^{\infty}\mathrm{d}s\eta(s)\delta(q(\xi-y+cs+c|y|/v)-h)q'(\xi-y+cs+c|y|/v). \tag{3.11}$$

Proceeding as above for the construction of the Evans function it is simple to show that

$$q'(\xi)=u(\xi)|_{\lambda=0}, \tag{3.12}$$

where $u=\mathcal{L}u$. Thus $q'(\xi)$ is an eigenfunction of $\mathcal{L}$ for $\lambda=0$ (expected from translation invariance). It is also possible to check that $\mathcal{E}(0)=0$, as expected. Introducing

$$\mathcal{H}(\lambda)=\int_0^{\infty}\mathrm{d}y w(y)\eta(y/c-y/v)\mathrm{e}^{-\lambda y/c} \tag{3.13}$$

and using the fact that $\mathcal{E}(0)=0$ allow us to write the Evans function in the form

$$\mathcal{E}(\lambda)=1-\frac{\mathcal{H}(\lambda)}{\mathcal{H}(0)}. \tag{3.14}$$

To calculate the essential spectrum of the operator $\mathcal{L}$ we make use of the fact that $u(\xi)$ has the convolution form

$$u(\xi)=\frac{u(0)}{c|q'(0)|}\int_{-\infty}^{\infty}\mathrm{d}y w(y)\eta(y/c_\pm-\xi/c)\mathrm{e}^{-\lambda(y-\xi)/c}. \tag{3.15}$$

Introducing the Fourier transform

$$\widehat{u}(k)=\int_{-\infty}^{\infty}u(\xi)\mathrm{e}^{-ikx}\mathrm{d}\xi, \tag{3.16}$$

taking the Fourier transform of (3.15), and rearranging and taking the inverse Fourier transform give

$$\frac{|q'(0)|}{u(0)}\int_{-\infty}^{\infty}\mathrm{d}k\frac{\widehat{u}(k)}{\widehat{\eta}(-i\lambda-kc)}\mathrm{e}^{ik\xi}=\int_{-\infty}^{\infty}\mathrm{d}k\widehat{w}\left(\frac{kc}{c_\pm}\pm i\frac{\lambda}{v}\right)\mathrm{e}^{ik\xi}. \tag{3.17}$$

Assuming that $w(\xi)$ decays exponentially quickly then in the limit $|\xi|\rightarrow\infty$ the right-hand side of (3.17) vanishes. Seeking solutions of the form $u(\xi)=\mathrm{e}^{ip\xi}$, where $p\in\mathbb{R}$, gives

$$\frac{1}{\widehat{\eta}(-i\lambda-pc)}=0. \tag{3.18}$$

This is an implicit equation for $\lambda=\lambda(p)$ that defines the essential spectrum of $\mathcal{L}$.

3.1. Example: A traveling front. Here we consider the choice $\eta(t) = \alpha e^{-\alpha t}$ and $w(x) = e^{-|x|}/2$ so that we recover a model previously analyzed by Zhang [41]. Assuming $c > 0$ the traveling front (2.7) is given in terms of (3.1) which takes the explicit form

$$\psi(\xi) = \begin{cases} \frac{1}{2}e^{m_-\xi}, & \xi \geq 0, \\ 1 - \frac{1}{2}e^{m_+\xi}, & \xi < 0, \end{cases} \qquad m_\pm = \frac{v}{c \pm v}. \tag{3.19}$$

The speed of the front is determined from the condition $q(0) = h$ as

$$c = \frac{v(2h-1)}{2h - 1 - 2hv/\alpha}. \tag{3.20}$$

The essential spectrum is easily found using

$$\widehat{\eta}(k) = \frac{1}{1 + ik/\alpha}, \tag{3.21}$$

so that, from (3.18), $\lambda(p) = -\alpha + ipc$, i.e., a vertical line at Re $\lambda = -\alpha$. Hence, as stated earlier, the real part of the continuous spectrum has a uniformly negative upper bound and is not important for determining wave stability. The Evans function is easily calculated using the result

$$\mathcal{H}(\lambda) = \frac{\alpha}{2} \frac{1}{1 + \alpha\left(\frac{1}{c} - \frac{1}{v}\right) + \frac{\lambda}{c}}, \tag{3.22}$$

so that

$$\mathcal{E}(\lambda) = \frac{\lambda}{c + \alpha\left(1 - \frac{c}{v}\right) + \lambda}. \tag{3.23}$$

The equation $\mathcal{E}(\lambda) = 0$ has only the solution $\lambda = 0$. We also have that

$$\mathcal{E}'(0) = \frac{1}{c + \alpha\left(1 - \frac{c}{v}\right)} > 0, \tag{3.24}$$

showing that $\lambda = 0$ is a simple eigenvalue. Hence, the traveling wave front for this example is exponentially stable. Note that in the limit $v \to \infty$ we recover the result of Zhang [41]. We make the observation that axonal communication delays affect wave speed but not wave stability.

4. Traveling waves in a model with linear recovery.

In real cortical tissues there is an abundance of metabolic processes whose combined effect is to modulate neuronal response. It is convenient to think of these processes in terms of local feedback mechanisms that modulate synaptic currents. Such feedback may act to decrease activity in the wake of a traveling front so as to generate traveling pulses (rather than fronts). We will consider simple models of so-called *spike frequency adaptation* (i.e., the addition of a current that activates in the presence of high activity) that are known to lead to the generation of pulses for network connectivities that would otherwise support only traveling fronts [34]. Generalizing the model in the previous section, we write

$$Qu(x,t) = (w \otimes f \circ u)(x,t) - ga(x,t), \tag{4.1}$$

$$Q_a a(x,t) = u(x,t), \tag{4.2}$$

and we have introduced the notation

$$(w \otimes f)(x,t) = \int_{-\infty}^{\infty} w(y) f(x-y, t-|y|/v) \mathrm{d}y. \tag{4.3}$$

The (temporal) linear differential operators Q and Q_a have Green's functions $\eta(t)$ and $\eta_a(t)$, respectively, so that

$$Q\eta(t) = \delta(t), \qquad Q_a \eta_a(t) = \delta(t). \tag{4.4}$$

This is a generalization of the model defined by (1.1) and (1.2). In its integrated form this more general model may be written

$$u = \eta * w \otimes f \circ u - g\eta_b * u, \tag{4.5}$$

where

$$(\eta * f)(x,t) = \int_0^t \eta(s) f(x, t-s) \mathrm{d}s \tag{4.6}$$

and $\eta_b = \eta * \eta_a$. Proceeding as before, we find that traveling wave solutions are given by

$$q(\xi) = \int_0^{\infty} \eta(s) \psi(\xi + cs) \mathrm{d}s - g \int_0^{\infty} \eta_b(s) q(\xi + cs) \mathrm{d}s, \tag{4.7}$$

with $\psi(\xi)$ given by (2.8). To obtain a solution for $q(\xi)$ it is convenient to take Fourier transforms and rearrange to give

$$\widehat{q}(k) = \widehat{\eta_c}(k) \widehat{\psi}(k), \qquad \widehat{\eta_c}(k) = \frac{\widehat{\eta}(k)}{1 + g\widehat{\eta_b}(k)}, \tag{4.8}$$

where $\widehat{\eta_b}(k) = \widehat{\eta}(k)\widehat{\eta_a}(k)$. Equation (4.8) may then be inverted to give $q(\xi)$ in the explicit form

$$q(\xi) = \int_0^{\infty} \eta_c(z) \psi(\xi + cz) \mathrm{d}z, \tag{4.9}$$

where

$$\eta_c(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\eta_c}(k) \mathrm{e}^{ikt} \mathrm{d}k. \tag{4.10}$$

This last integral may be evaluated using a contour in the upper half complex plane for $t > 0$ (with $\eta_c(t) = 0$ for $t < 0$).

A traveling front solution is given by (4.9) and (3.1). The speed of the wave is again determined by the condition $q(0) = h$. Linearizing around the traveling front and proceeding as before (for the case without recovery), we obtain an eigenvalue equation of the form $u = \mathcal{L}_c u$, where

$$\mathcal{L}_c u = \frac{u(0)}{c|q'(0)|} \int_{-\infty}^{\infty} \mathrm{d}y w(y) \eta_c(-\xi/c + y/c - |y|/v) \mathrm{e}^{-\lambda(y-\xi)/c}. \tag{4.11}$$

The essential spectrum $\lambda = \lambda(p)$ is determined from the solution of $1/\widehat{\eta}_c(-i\lambda - pc) = 0$. The Evans function of a front in this model with recovery is given by (3.14) with

$$\mathcal{H}(\lambda) = \int_0^{\infty} \mathrm{d}y w(y) \eta_c(y/c - y/v) \mathrm{e}^{-\lambda y/c}. \tag{4.12}$$

The model defined by (4.5) is also expected to support traveling pulses of the form $q(\xi) \geq h$ for $\xi \in [0, \Delta]$ and $q(\xi) < h$ otherwise. In this case the expression for $\psi(\xi)$ is given by

$$\psi(\xi) = \begin{cases} \mathcal{F}\left(\frac{-\xi}{1+c/v}, \frac{\Delta-\xi}{1+c/v}\right), & \xi \leq 0, \\ \mathcal{F}\left(0, \frac{\xi}{1-c/v}\right) + \mathcal{F}\left(0, \frac{\Delta-\xi}{1+c/v}\right), & 0 < \xi < \Delta, \\ \mathcal{F}\left(\frac{\xi-\Delta}{1-c/v}, \frac{\xi}{1-c/v}\right), & \xi \geq \Delta, \end{cases} \tag{4.13}$$

where

$$\mathcal{F}(a, b) = \int_a^b w(y) \mathrm{d}y. \tag{4.14}$$

The dispersion relation $c = c(\Delta)$ is then implicitly defined by the simultaneous solution of $q(0) = h$ and $q(\Delta) = h$ $(\Delta > 0)$. Linearizing around the traveling pulse solution and proceeding as before, we obtain an eigenvalue equation of the form $u = \mathcal{J}_c u$, where for $\xi \in [0, \Delta]$

$$\mathcal{J}_c u(\xi) = A_c(\xi, \lambda) u(0) + B_c(\xi, \lambda) u(\Delta). \tag{4.15}$$

Here the functions $A_c(\xi, \lambda)$ and $B_c(\xi, \lambda)$ are given by

$$A_c(\xi, \lambda) = \frac{1}{c|q'(0)|} \int_{\frac{\xi}{1-c/v}}^{\infty} \mathrm{d}y w(y) \eta_c(-\xi/c + y/c - y/v) \mathrm{e}^{-\lambda(y-\xi)/c}, \tag{4.16}$$

$$B_c(\xi, \lambda) = \frac{1}{c|q'(\Delta)|} \int_{\frac{\xi-\Delta}{1+c/v}}^{\infty} \mathrm{d}y w(y) \eta_c((\Delta - \xi)/c + y/c - |y|/v) \mathrm{e}^{-\lambda(y-(\xi-\Delta))/c}. \tag{4.17}$$

Demanding that the eigenvalue problem $u = \mathcal{J}_c u$ be self-consistent at $\xi = 0$ and $\xi = \Delta$ gives the system of equations

$$\begin{bmatrix} u(0) \\ u(\Delta) \end{bmatrix} = \mathcal{A}_c(\lambda) \begin{bmatrix} u(0) \\ u(\Delta) \end{bmatrix}, \qquad \mathcal{A}_c(\lambda) = \begin{bmatrix} A_c(0, \lambda) & B_c(0, \lambda) \\ A_c(\Delta, \lambda) & B_c(\Delta, \lambda) \end{bmatrix}. \tag{4.18}$$

There is a nontrivial solution of (4.18) if $\mathcal{E}(\lambda)=0$, where $\mathcal{E}(\lambda)=\det(\mathcal{A}_c(\lambda)-I)$. We interpret $\mathcal{E}(\lambda)$ as the Evans function for the traveling pulse solution.

Note that standing pulses are defined by $c=0$ so that from (4.9)

$$q(\xi)=\widehat{\eta}_c(0)\int_0^\Delta w(\xi-y)\mathrm{d}y. \tag{4.19}$$

Hence, for a standing pulse

$$q'(\xi)=\widehat{\eta}_c(0)[w(\xi)-w(\xi-\Delta)], \tag{4.20}$$

and $|q'(0)|=|q'(\Delta)|$. Moreover, since $w(y)$ is relatively flat compared to $\eta_c(y/c)\mathrm{e}^{-\lambda y/c}/c$ when $c=0$, the expressions for (4.16) and (4.17) simplify to

$$A_c(\xi,\lambda)=\frac{1}{|q'(0)|}\widehat{\eta}_c(-i\lambda)w(\xi)\mathrm{e}^{-\lambda\xi/v}, \qquad B_c(\xi,\lambda)=A_c(\Delta-\xi,\lambda). \tag{4.21}$$

In this case it can be shown that the Evans function $\mathcal{E}(\lambda)$ has zeros when

$$\frac{\widehat{\eta}_c(0)}{\widehat{\eta}_c(-i\lambda)}=\Gamma_\pm(\lambda), \tag{4.22}$$

where

$$\Gamma_\pm(\lambda)=\frac{w(0)\pm w(\Delta)\mathrm{e}^{-\lambda\Delta/v}}{|w(0)-w(\Delta)|}. \tag{4.23}$$

Note that $\lambda=0$ is a solution as expected.

4.1. Example: A front bifurcation.

Here we consider an example that recovers a model recently discussed by Bressloff and Folias [6] by choosing $\eta(t)=\alpha\mathrm{e}^{-\alpha t}$, $\eta_a(t)=\mathrm{e}^{-t}$, and $w(x)=\mathrm{e}^{-|x|}/2$. From (4.10) the function $\eta_c(t)$ is easily calculated given the pole structure of $\widehat{\eta}_c(k)$. Using the facts that $\widehat{\eta}(k)=(1+ik/\alpha)^{-1}$ and $\widehat{\eta}_a(k)=(1+ik)^{-1}$, this is given by

$$\widehat{\eta}_c(k)=\frac{-\alpha(1+ik)}{(k-ik_+)(k-ik_-)}, \tag{4.24}$$

where

$$k_\pm=\frac{1+\alpha\pm\sqrt{(1+\alpha)^2-4\alpha(1+g)}}{2}. \tag{4.25}$$

Hence, upon evaluating (4.10), using the calculus of residues, we obtain

$$\eta_c(t)=\frac{\alpha}{k_--k_+}\left\{(1-k_+)\mathrm{e}^{-k_+t}-(1-k_-)\mathrm{e}^{-k_-t}\right\}. \tag{4.26}$$

Using (4.9) and (3.19), the equation $q(0)=h$ gives an implicit expression for the front speed as

$$h=\frac{\alpha}{2}\frac{1-cm_-}{(cm_-)^2-cm_-(1+\alpha)+\alpha(1+g)}, \qquad c>0, \tag{4.27}$$

$$h=-\frac{\alpha}{2}\frac{1-cm_+}{(cm_+)^2-cm_+(1+\alpha)+\alpha(1+g)}+\frac{1}{1+g}, \qquad c<0. \tag{4.28}$$

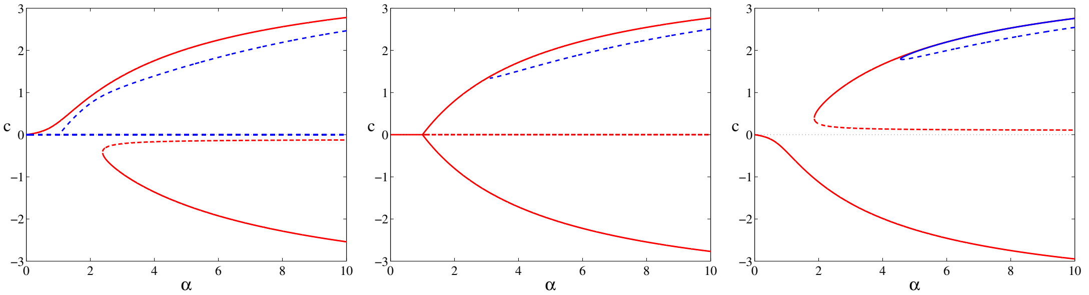


Figure 4.1. *Wave front speed as a function of α for a model with linear recovery (red curve). If $g = g_c$, where $2h(1+g_c) = 1$, there is a front for all α with speed $c = 0$. At a critical value of α this stationary front undergoes a pitchfork bifurcation leading to a pair of fronts traveling in opposite directions. This case is illustrated in the middle figure, where $v = 4$, $h = 0.25$, and $g = 1$. Away from this critical condition the pitchfork bifurcation is broken as illustrated in the left and right figures, where $g = 0.9$ and $g = 1.1$, respectively. Solid (dashed) lines are stable (unstable). The blue curves illustrate the existence of stable (solid) and unstable (dashed) pulses, whose speed and width are determined by the simultaneous solution of* (4.33) *and* (4.34).

Note that in the limit $v \to \infty$, $m_\pm \to \pm 1$ and we recover the result of Bressloff and Folias [6]. Rearranging (4.27) and (4.28) gives c in the form

$$(4.29)\qquad cm_- = \frac{1}{2}\left[1+\alpha-\frac{\alpha}{2h} \pm \sqrt{\left(1+\alpha-\frac{\alpha}{2h}\right)^2 - 4\alpha\left(1+g-\frac{1}{2h}\right)}\right], \qquad c > 0,$$

$$(4.30)\qquad cm_+ = \frac{1}{2}\left[1+\alpha-\frac{\alpha}{2h^*} \pm \sqrt{\left(1+\alpha-\frac{\alpha}{2h^*}\right)^2 - 4\alpha\left(1+g-\frac{1}{2h^*}\right)}\right], \qquad c < 0,$$

where $h^* = 1/(1+g) - h$. If $g = g_c$, where $2h(1+g_c) = 1$, there is a front for all α with speed $c = 0$. At a critical value of α this stationary front undergoes a pitchfork bifurcation leading to a pair of fronts traveling in opposite directions. If this critical condition is not met, then the pitchfork bifurcation is broken as illustrated in Figure 4.1. Since the zeros of $1/\widehat{\eta_c}(k) = 0$ occur when $k = ik_\pm$, we see that the essential spectrum is contained within the closed strip bounded by the two vertical lines $\lambda(p) = -k_\pm + ipc$. The Evans function may also be easily computed using

$$(4.31)\qquad \mathcal{H}(\lambda) = \frac{\alpha}{2(k_- - k_+)}\left\{\frac{1-k_+}{1+k_+(\frac{1}{c}-\frac{1}{v})+\frac{\lambda}{c}} - \frac{1-k_-}{1+k_-(\frac{1}{c}-\frac{1}{v})+\frac{\lambda}{c}}\right\}.$$

On the branch with $c = 0$ and $g = g_c$ (defining a stationary front) we find that

$$(4.32)\qquad \mathcal{E}(\lambda) = \lambda\frac{(\lambda + k_+ + k_- - k_+k_-)}{(\lambda+k_+)(\lambda+k_-)},$$

which has zeros when $\lambda = 0$ and $\lambda = k_+k_- - (k_+ + k_-) = \alpha g_c - 1$. Hence, the stationary front changes from stable to unstable as α is increased through $1/g_c$. This result may also be used to infer the stability of the other branches in Figure 4.1 (rather than laboriously evaluating the Evans function on each branch).

Examples of stationary fronts and fronts traveling in opposite directions, as predicted by the above analysis, are illustrated in Figure 4.2. These simulations were implemented using the numerical scheme of Hutt, Bestehorn, and Wennekers [22] to calculate the integral arising on the right-hand side of (2.1). Fourier methods, as used previously by Coombes, Lord, and Owen [13], were found to give similar results. In Figure 4.2, $g = 1.0, h = 0.25$, and $v = 4$, as in the second part of Figure 4.1. The stationary front is stable for $\alpha = 0.9$, but for $\alpha > 1$ it loses stability in favor of a pair of moving fronts. As illustrated in Figure 4.3, this symmetry breaking is reflected in the u-a phase-plane. Here the stationary front lies on $a = u$, while the forward and backward fronts are distinguished by their trajectories on either side of $a = u$. Similar waves and bifurcations are found in a wide class of reaction-diffusion systems [20]. Note that initial conditions must be specified along with a history: the forward wave is generated by stimulating a previously inactive region, whereas the reverse wave is initiated by removing stimulation from a previously active region.

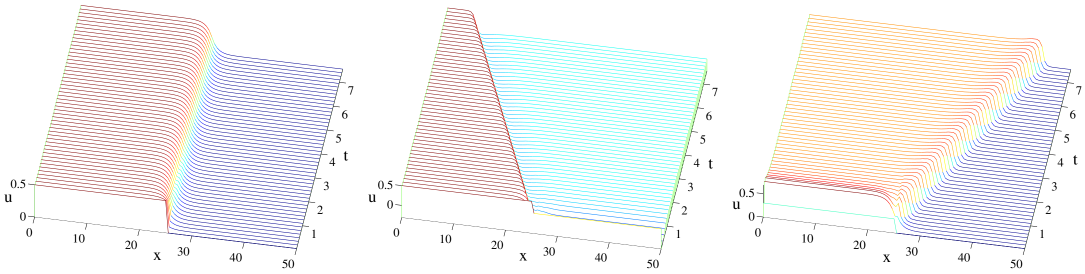


Figure 4.2. *Illustrations of stationary, backward, and forward waves. In all cases, $g = 1.0$, $h = 0.25$, and $v = 4$, corresponding to the second part of Figure 4.1. The stationary front has $\alpha = 0.9$, and the propagating waves have $\alpha = 10$.*

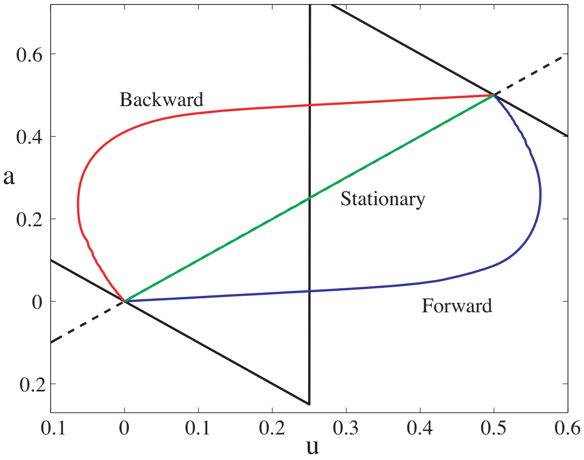
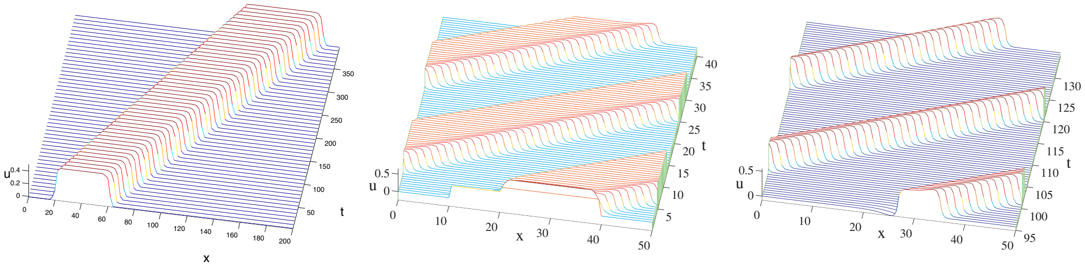


Figure 4.3. *Phase planes corresponding to Figures 4.1 and 4.2. The dashed black line denotes $a = u$, and the solid black line $a = (\Theta(u - h) - u)/g$. Forward and backward waves connect the same fixed points, but the different trajectories describe propagation in opposite directions.*

4.1.1. Beyond the front bifurcation. Motivated by results from the last example, it is interesting to consider whether a stable traveling front and back with different speeds (that are stable in isolation) can interact to form new types of persistent solutions. Numerical simulations show that initially well-separated front and back solutions move apart if the relative speed of the two is positive and converge if negative. In the former case this leads to a widening region of activity of the type shown in Figure 4.4(a). In the latter case the fronts can either annihilate or merge to form a traveling pulse, as illustrated in Figures 4.4(b) and 4.4(c).



(a) (b) (c)

Figure 4.4. *Interacting fronts either separate or converge. In the latter case a stable pulse may arise if parameters allow.* (a) $g = 0.9$*;* (b), (c) $g = 1.1$*; the front moves more slowly than the back, and a pulse exists for these parameters (*$v = 4, \alpha = 10, h = 0.25$*). Part* (b) *shows the early evolution as the back begins to catch up with the front, and part* (c) *shows the final convergence to a pulse solution. Clicking on the above images displays the associated movies* (60595_01.mov *and* 60595_02.mov).

The parameter regime that supports traveling pulses can be found by explicit construction. The existence of a pulse solution is determined by enforcing $q(0) = h = q(\Delta)$, giving

(4.33)
$$h = \frac{\alpha(1 - e^{\Delta m_-})(1 - cm_-)}{2((cm_-)^2 - cm_-(1+\alpha) + \alpha(1+g))},$$

$$h = \frac{1}{1+g} + \alpha\frac{cm_- - 1}{2((cm_-)^2 - cm_-(1+\alpha) + \alpha(1+g))} - \alpha\frac{(1 - cm_+)e^{-\Delta m_+}}{2((cm_+)^2 - cm_+(1+\alpha) + \alpha(1+g))}$$
$$+ \frac{\alpha}{k_- - k_+}\left\{(1 - k_-)e^{-\Delta k_-/c}\left(\frac{1}{k_-} + \frac{1}{2(cm_- - k_-)} + \frac{1}{2(cm_+ - k_-)}\right)\right.$$

(4.34)
$$\left. -(1 - k_+)e^{-\Delta k_+/c}\left(\frac{1}{k_+} + \frac{1}{2(cm_- - k_+)} + \frac{1}{2(cm_+ - k_+)}\right)\right\}.$$

Figure 4.5 shows the speed and width of such a traveling pulse as the conduction velocity v varies. Note that pulses are not supported when the conduction velocity is too small, as there is a fold bifurcation with decreasing v. The speed of pulses as a function of α is also plotted in conjunction with that of fronts in Figure 4.1, showing that stable pulses are preferred for $g > g_c$ and sufficiently large α.

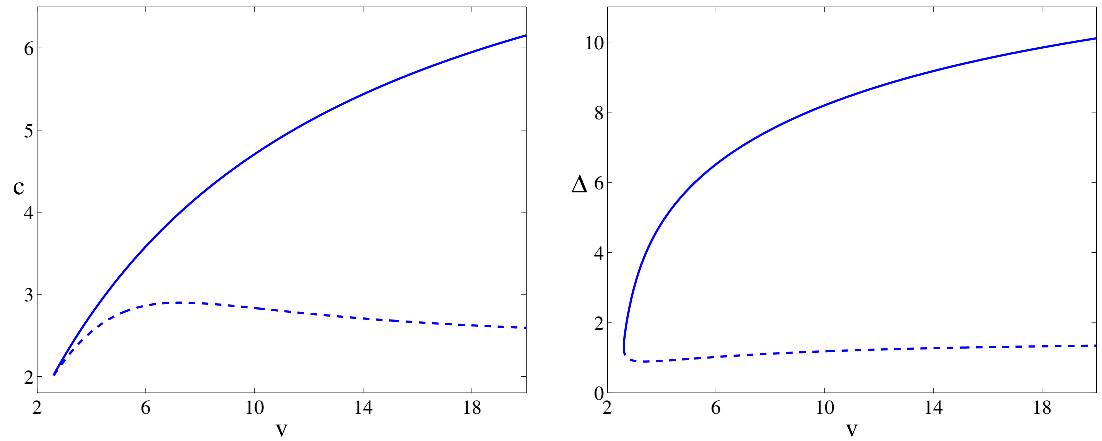


Figure 4.5. *Speed and width of the traveling pulse as a function of v in a neural model with linear recovery. The faster wave has the largest width. We note the fast and slow traveling pulse are annihilated in a saddle-node bifurcation with decreasing v. Here, $h = 0.25, \alpha = 1$, and $g = 1.1$. Solid (dashed) lines are stable (unstable).*

4.2. Example: An unstable standing pulse. In this example we consider the system discussed in section 4.1 but focus on standing pulses rather than traveling waves. For simplicity we shall also consider the limit $v \to \infty$. From (4.13)

$$\psi(\xi) = \begin{cases} \frac{1}{2}(\mathrm{e}^{\xi} - \mathrm{e}^{\xi-\Delta}), & \xi \leq 0, \\ 1 - \frac{1}{2}(\mathrm{e}^{\xi-\Delta} + \mathrm{e}^{-\xi}), & 0 < \xi < \Delta, \\ \frac{1}{2}(\mathrm{e}^{-(\xi-\Delta)} - \mathrm{e}^{-\xi}), & \xi \geq \Delta, \end{cases} \tag{4.35}$$

and from (4.8) $\widehat{\eta}_c(0) = 1/(1+g)$. The pulse width is determined by setting $q(0) = h$ or, equivalently, $q(\Delta) = h$ to give

$$\Delta = -\ln[1 - 2h(1+g)], \qquad h \leq \frac{1}{2(1+g)}. \tag{4.36}$$

From (4.22) and (4.8) the zeros of the Evans function satisfy

$$\lambda^2 + \lambda[1 + \alpha - \alpha(1+g)\Gamma_\pm(0)] - \alpha(1+g)(\Gamma_\pm(0) - 1) = 0. \tag{4.37}$$

Since $\Gamma_-(0) = 1$ there are solutions of (4.37) with $\lambda = 0$ and $\lambda = \alpha g - 1$. The remaining solutions are given by

$$\lambda_\pm = \frac{-\Lambda \pm \sqrt{\Lambda + 4\alpha(1+g)(\Gamma_+(0) - 1)}}{2}, \tag{4.38}$$

with

$$\Lambda = 1 + \alpha - \alpha(1+g)\Gamma_+(0). \tag{4.39}$$

Since $\Gamma_+(0) > 1$ it follows that $\lambda_+ > 0$, and, hence, the stationary pulse is always unstable. In section 5 we shall consider a model with nonlinear recovery that can support a stable standing pulse.

4.3. Example: A pair of traveling pulses. Once again we consider the system discussed in section 4.2 but construct traveling rather than standing pulses. However, to recover a model discussed by Pinto and Ermentrout [34] we consider the case $Q_a = \partial_t$ or, equivalently, $\widehat{\eta}_a(k) = \lim_{\epsilon \to 0}(\epsilon + ik)^{-1}$. The function $\eta_c(t)$ is easily calculated as

$$\eta_c(t) = \frac{\alpha}{k_+ - k_-}\left\{k_+ \mathrm{e}^{-k_+ t} - k_- \mathrm{e}^{-k_- t}\right\}, \tag{4.40}$$

with

$$k_\pm = \frac{\alpha \pm \sqrt{\alpha^2 - 4\alpha g}}{2} \tag{4.41}$$

and $\psi(\xi)$ given by (4.35). It just remains to enforce the conditions $q(0) = h = q(\Delta)$, giving the two equations

$$\begin{aligned} h &= \frac{\alpha c(1 - \mathrm{e}^{-\Delta})}{2(c^2 + \alpha c + \alpha g)}, \\ h &= \frac{\alpha}{k_+ - k_-}\left\{\mathrm{e}^{-k_- \Delta/c} + \frac{k_-}{2}\left(\frac{\mathrm{e}^{-\Delta} - \mathrm{e}^{-k_- \Delta/c}}{k_- - c} + \frac{1 - \mathrm{e}^{-k_- \Delta/c}}{k_- + c}\right)\right. \\ &\qquad \left. -\mathrm{e}^{-k_+ \Delta/c} - \frac{k_+}{2}\left(\frac{\mathrm{e}^{-\Delta} - \mathrm{e}^{-k_+ \Delta/c}}{k_+ - c} + \frac{1 - \mathrm{e}^{-k_+ \Delta/c}}{k_+ + c}\right)\right\}. \end{aligned} \tag{4.42}$$

We plot the simultaneous solution of these two equations in Figure 4.6, showing the speed and width of the traveling pulse as a function of g. We note that there are two solution branches: one describing a fast wide pulse and the other a slower narrower pulse. Motivated by numerical experiments, Pinto and Ermentrout have conjectured that the larger (fast) pulse is stable and the narrower (slower) pulse unstable. We are now in a position to confirm this

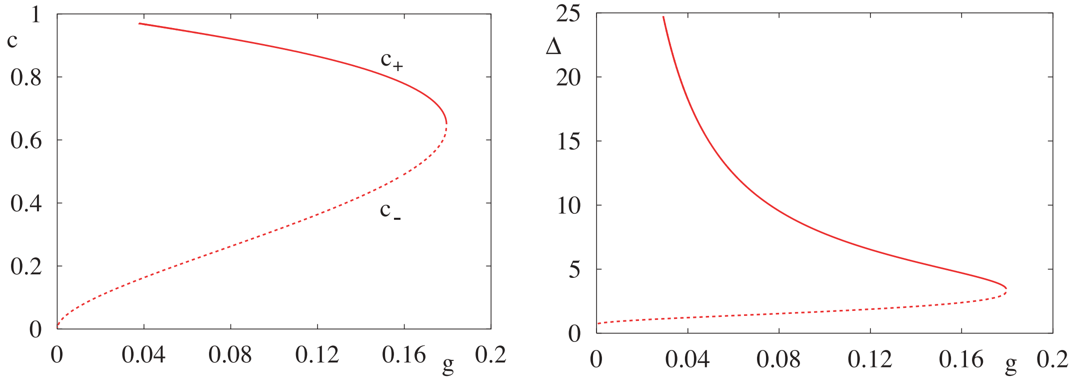


Figure 4.6. *Speed and width of the traveling pulse as a function of g in a neural model with linear recovery. The faster wave, with speed c_+, has the largest width. We note that the fast and slow traveling pulse are annihilated in a saddle-node bifurcation with increasing g and that the width of the faster pulse diverges to infinity with decreasing g. Here, $h = 0.25$ and $\alpha = 1$. Solid (dashed) lines are stable (unstable).*

by examining the Evans function for a traveling pulse. The functions $A_c(0,\lambda)$ and $B_c(0,\lambda)$ are calculated to be

$$A_c(0,\lambda) = \frac{1}{c|q'(0)|}\frac{\alpha}{k_+ - k_-}\frac{1}{2}\left\{\frac{k_+}{1+k_+/c+\lambda/c} - \frac{k_-}{1+k_-/c+\lambda/c}\right\}, \tag{4.43}$$

$$\begin{aligned} B_c(0,\lambda) = \frac{1}{c|q'(\Delta)|}\frac{\alpha}{k_+ - k_-}\frac{1}{2}\Bigg\{ & k_+\left(\frac{\mathrm{e}^{-(k_+ + \lambda)\Delta/c} - \mathrm{e}^{-\Delta}}{1-k_+/c-\lambda/c} + \frac{\mathrm{e}^{-(k_+ + \lambda)\Delta/c}}{1+k_+/c+\lambda/c}\right) \\ & - k_-\left(\frac{\mathrm{e}^{-(k_- + \lambda)\Delta/c} - \mathrm{e}^{-\Delta}}{1-k_-/c-\lambda/c} + \frac{\mathrm{e}^{-(k_- + \lambda)\Delta/c}}{1+k_-/c+\lambda/c}\right)\Bigg\}, \end{aligned} \tag{4.44}$$

with $A_c(\Delta,\lambda) = \mathrm{e}^{-\Delta}A_c(0,\lambda)$, $B_c(\Delta,\lambda) = |q'(0)/q'(\Delta)|A_c(0,\lambda)$. Using the fact that for $v \to \infty$,

$$q'(\xi) = \int_0^\infty \eta_c(s)[w(\xi+cs) - w(\xi-\Delta+cs)]\mathrm{d}s, \tag{4.45}$$

we have that $q'(\Delta) = -h$ and

$$\begin{aligned} q'(0) = \frac{h}{1-\mathrm{e}^{-\Delta}} - \frac{\alpha}{2(k_+ - k_-)}\Bigg\{ & \frac{k_+(\mathrm{e}^{-k_+\Delta/c} - \mathrm{e}^{-\Delta})}{c-k_+} - \frac{k_-(\mathrm{e}^{-k_-\Delta/c} - \mathrm{e}^{-\Delta})}{c-k_-} \\ & + \frac{k_+\mathrm{e}^{-k_+\Delta/c}}{c+k_+} - \frac{k_-\mathrm{e}^{-k_-\Delta/c}}{c+k_-}\Bigg\}. \end{aligned} \tag{4.46}$$

The Evans function $\mathcal{E}(\lambda) = \det(\mathcal{A}_c(\lambda) - I)$ may then be calculated using (4.18). In Figure 4.7 we show a section of the Evans function along the real axis for a wave on the fast branch and a wave on the slow branch. Also plotted is the Evans function for the wave that arises at the limit point in Figure 4.6, where the fast and slow waves annihilate. This figure nicely illustrates that for a wave on the fast branch the Evans function has no zeros on the positive real axis, while the slow wave does. Moreover, as one moves around the branch from a fast to a slow wave the Evans function develops a repeated root at the origin, as expected. To further illustrate that the traveling pulse changes from stable to unstable as one moves around the limit point in Figure 4.6, we track out the zero solution from Figure 4.7 along the solution branch of Figure 4.6 and plot this in Figure 4.8.

5. Pulses in a model of nonlinear recovery.

In this section we consider a system in which the recovery variable is governed by a nonlinear model, rather than a linear one as in section 4. Moreover, we shall consider the recovery process itself to be nonlocal and write

$$Qu(x,t) = (w \otimes f \circ u)(x,t) - g(w_a \otimes a)(x,t), \tag{5.1}$$

$$Q_a a(x,t) = f \circ u(x,t). \tag{5.2}$$

In keeping with earlier sections we consider the case when $w_a(x) = w_a(|x|)$. In its integrated form this model may be written

$$u = [\eta * w \otimes - g\eta_b * w_a \otimes] f \circ u. \tag{5.3}$$

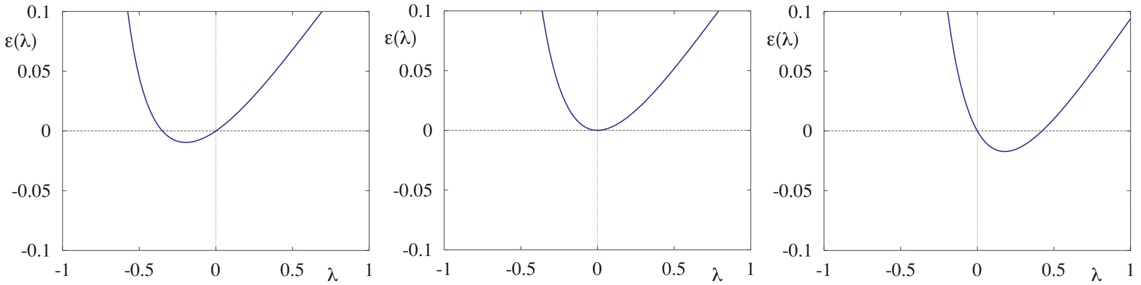


Figure 4.7. *A plot of $\mathcal{E}(\lambda)$ along the real axis for three different points on the solution branch shown in Figure 4.6. In the figure on the left $g = 0.15$ with $c = c_+$ showing that a point on the fast branch has at least one zero on the negative real axis. Plotting over a much wider domain shows that there are no zeros on the positive real axis and that this graph asymptotes to 1. In the middle figure we see a repeated root at $\lambda = 0$ when $g = .1793$, corresponding to the limit point in Figure 4.6 where a fast and a slow wave merge. In the right-hand figure $g = 0.15$ with $c = c_-$, and there is a zero of the Evans function on the positive real axis, showing that the slow branch is unstable. Note that $\lambda = 0$ is always a solution (as expected from translation invariance).*

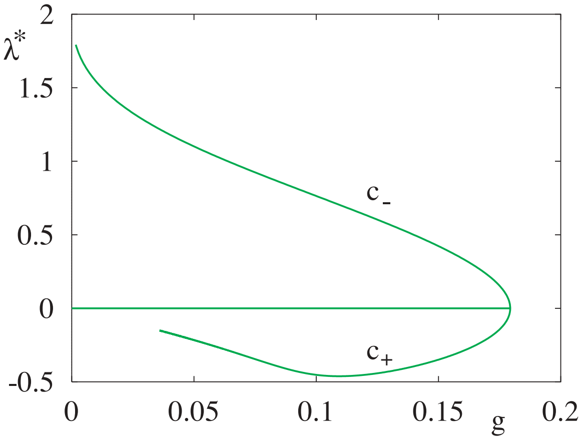


Figure 4.8. *A plot of the zero of the Evans function, denoted λ^*, as seen in Figure 4.7 along the solution branch seen in Figure 4.6. Note that there is always a branch of solutions with $\lambda^* = 0$. The other branch passes through the origin, where the slow and fast waves of Figure 4.6 merge. We conclude that the fast wave ($c = c_+$) is stable and the slow wave ($c = c_-$) is unstable.*

This can be interpreted as a lateral inhibitory network model as in the paper of Pinto and Ermentrout, or for the case with a relatively narrow footprint for $w_a(x)$ it may be regarded as a variant of the model with recovery described in section 4. In either case we shall show that the model can support stable standing pulses, unlike the case when the recovery variable evolves according to a linear model.

In a comoving frame we have a modified form of (2.2) under the replacement $w(y)\eta(s) \rightarrow$

$w(y)\eta(s) - gw_a(y)\eta_b(s)$. Hence, traveling wave solutions are given by

$$q(\xi) = \left(\int_{-\infty}^{\infty} \mathrm{d}y w(y) \int_0^{\infty} \mathrm{d}s\eta(s) - g\int_{-\infty}^{\infty} \mathrm{d}y w_a(y) \int_0^{\infty} \mathrm{d}s\eta_b(s)\right) \times \Theta(q(\xi - y + cs + c|y|/v) - h). \tag{5.4}$$

Linearizing around a traveling pulse solution and proceeding as before, we obtain an eigenvalue equation of the form $u = \mathcal{L}u - g\mathcal{J}u$, where

$$\mathcal{J}u = \int_{-\infty}^{\infty} \mathrm{d}y w_a(y)\mathrm{e}^{-\lambda|y|/v} \int_0^{\infty} \mathrm{d}s\eta_b(s)\mathrm{e}^{-\lambda s} f'(q(\xi - y + cs + c|y|/v)) \times u(\xi - y + cs + c|y|/v), \tag{5.5}$$

and $\mathcal{L}$ is defined by (2.5). We may then proceed analogously to the case for the front solution described in section 3, for $\xi \in [0, \Delta]$, to obtain

$$\mathcal{L}u(\xi) = A(\xi,\lambda)u(0) + B(\xi,\lambda)u(\Delta), \qquad \mathcal{J}u(\xi) = C(\xi,\lambda)u(0) + D(\xi,\lambda)u(\Delta), \tag{5.6}$$

where

$$A(\xi,\lambda) = \frac{1}{c|q'(0)|} \int_{\frac{\xi}{1-c/v}}^{\infty} \mathrm{d}y w(y)\eta(-\xi/c + y/c - y/v)\mathrm{e}^{-\lambda(y-\xi)/c}, \tag{5.7}$$

$$B(\xi,\lambda) = \frac{1}{c|q'(\Delta)|} \int_{\frac{\xi-\Delta}{1+c/v}}^{\infty} \mathrm{d}y w(y)\eta((\Delta - \xi)/c + y/c - |y|/v)\mathrm{e}^{-\lambda(y-(\xi-\Delta))/c}, \tag{5.8}$$

$$C(\xi,\lambda) = \frac{1}{c|q'(0)|} \int_{\frac{\xi}{1-c/v}}^{\infty} \mathrm{d}y w_a(y)\eta_b(-\xi/c + y/c - y/v)\mathrm{e}^{-\lambda(y-\xi)/c}, \tag{5.9}$$

$$D(\xi,\lambda) = \frac{1}{c|q'(\Delta)|} \int_{\frac{\xi-\Delta}{1+c/v}}^{\infty} \mathrm{d}y w_a(y)\eta_b((\Delta - \xi)/c + y/c - |y|/v)\mathrm{e}^{-\lambda(y-(\xi-\Delta))/c}. \tag{5.10}$$

Demanding that perturbations be determined self-consistently at $\xi = 0$ and $\xi = \Delta$ gives the system of equations

$$\begin{bmatrix} u(0) \\ u(\Delta) \end{bmatrix} = \mathcal{A}(\lambda) \begin{bmatrix} u(0) \\ u(\Delta) \end{bmatrix}, \qquad \mathcal{A}(\lambda) = \begin{bmatrix} A(0,\lambda) - gC(0,\lambda) & B(0,\lambda) - gD(0,\lambda) \\ A(\Delta,\lambda) - gC(\Delta,\lambda) & B(\Delta,\lambda) - gD(\Delta,\lambda) \end{bmatrix}. \tag{5.11}$$

There is a nontrivial solution of (5.11) if $\mathcal{E}(\lambda) = 0$, where $\mathcal{E}(\lambda) = \det(\mathcal{A}(\lambda) - I)$. We interpret $\mathcal{E}(\lambda)$ as the Evans function of a traveling pulse solution of (5.3). Working along identical lines to earlier sections the essential spectrum is defined by

$$\frac{1}{\widehat{\eta}(-i\lambda - pc)} \frac{1}{\widehat{\eta_a}(-i\lambda - pc)} = 0. \tag{5.12}$$

For standing pulses with $c = 0$, $Qu = u$ and $Q_a a = a$ so that

$$q(\xi) = \int_0^{\Delta} w_b(\xi - y)\mathrm{d}y, \tag{5.13}$$

where we have introduced the effective interaction kernel $w_b(x) = w(x) - gw_a(x)$. Hence,

$$q'(\xi) = w_b(\xi) - w_b(\xi - \Delta), \tag{5.14}$$

and we note that $|q'(0)| = |q'(\Delta)|$. For $c = 0$, $w(y)$ and $w_a(y)$ are relatively flat compared to $\eta(y/c)\mathrm{e}^{-\lambda y/c}/c$ and $\eta_b(y/c)\mathrm{e}^{-\lambda y/c}/c$, and we obtain the further simplification

$$A(\xi,\lambda) = \frac{1}{|q'(0)|}\widehat{\eta}(-i\lambda)w(\xi)\mathrm{e}^{-\lambda\xi/v}, \qquad B(\xi,\lambda) = A(\Delta - \xi, \lambda), \tag{5.15}$$

$$C(\xi,\lambda) = \frac{1}{|q'(0)|}\widehat{\eta}_b(-i\lambda)w_a(\xi)\mathrm{e}^{-\lambda\xi/v}, \qquad D(\xi,\lambda) = C(\Delta - \xi, \lambda). \tag{5.16}$$

5.1. Example: A pair of traveling pulses. Here we consider the choice $\eta(t) = \alpha\mathrm{e}^{-\alpha t}$, $\eta_a(t) = \mathrm{e}^{-t}$, $w(x) = \mathrm{e}^{-|x|}/2$, and $w_a = \delta(x)$ so that we recover a model recently discussed by Coombes, Lord, and Owen [13]. With the use of the Evans function we will now establish the earlier conjecture of these authors that of the two possible coexisting traveling pulses in this model, it is the faster of the two that is stable. The traveling pulse solution for this model is given by [13]

$$q(\xi) = \frac{\alpha}{c}\int_\xi^\infty \mathrm{e}^{\alpha(\xi - z)}[\psi(z) - ga(z)]\mathrm{d}z, \tag{5.17}$$

where

$$a(\xi) = \begin{cases} [1 - \mathrm{e}^{-\Delta/c}]\mathrm{e}^{\xi/c}, & \xi \le 0, \\ [1 - \mathrm{e}^{(\xi-\Delta)/c}], & 0 < \xi < \Delta, \\ 0, & \xi \ge \Delta. \end{cases} \tag{5.18}$$

Using (4.13) $\psi(\xi)$ is given by

$$\psi(\xi) = \begin{cases} \frac{1}{2}(\mathrm{e}^{m_+\xi} - \mathrm{e}^{m_+(\xi-\Delta)}), & \xi \le 0, \\ 1 - \frac{1}{2}(\mathrm{e}^{m_+(\xi-\Delta)} + \mathrm{e}^{m_-\xi}), & 0 < \xi < \Delta, \\ \frac{1}{2}(\mathrm{e}^{m_-(\xi-\Delta)} - \mathrm{e}^{m_-\xi}), & \xi \ge \Delta. \end{cases} \tag{5.19}$$

In Figure 5.1 we plot the speed of the pulse as a function of g, obtained by the simultaneous solution of $q(0) = h$ and $q(\Delta) = h$. This is reminiscent of that obtained for the model with linear recovery in section 4.3 (as also is the plot of $\Delta = \Delta(g)$, not shown). The essential spectrum for this problem is easily calculated and can be shown to contain the closed strip bounded by the two vertical lines $\lambda = -\alpha + ipc$ and $\lambda = -1 + ipc$. It is also straightforward

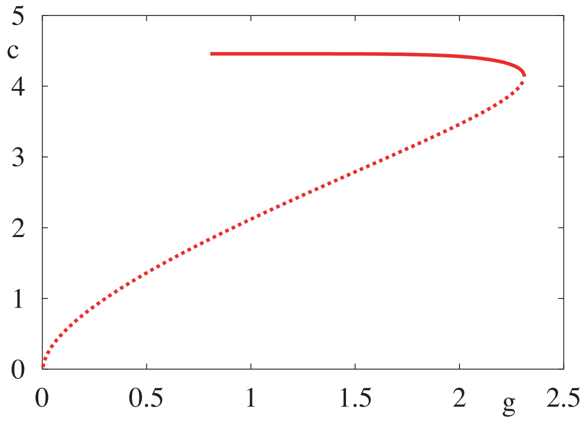


Figure 5.1. *Speed of traveling pulse as a function of g in a model with nonlinear recovery. Parameters are $h = 0.1$, $\alpha = 2$, and $v = 10$.*

to obtain $C(0,\lambda) = C(\Delta,\lambda) = D(\Delta,\lambda) = 0$ and

$$A(0,\lambda) = \frac{1}{c|q'(0)|}\frac{\alpha}{2}\frac{1}{1+\alpha\left(\frac{1}{c}-\frac{1}{v}\right)+\frac{\lambda}{c}}, \tag{5.20}$$

$$B(\Delta,\lambda) = \left|\frac{q'(0)}{q'(\Delta)}\right| A(0,\lambda), \tag{5.21}$$

$$A(\Delta,\lambda) = \mathrm{e}^{-\Delta(v+\lambda)/(v-c)}A(0,\lambda), \tag{5.22}$$

$$B(0,\lambda) = \frac{1}{c|q'(\Delta)|}\frac{\alpha}{2}\left\{\frac{\mathrm{e}^{-(\alpha+\lambda)\Delta/c}-\mathrm{e}^{-(v+\lambda)\Delta/(v+c)}}{1-\alpha(\frac{1}{c}+\frac{1}{v})-\frac{\lambda}{c}}+\frac{\mathrm{e}^{-(\alpha+\lambda)\Delta/c}}{1+\alpha\left(\frac{1}{c}-\frac{1}{v}\right)+\frac{\lambda}{c}}\right\}, \tag{5.23}$$

$$D(0,\lambda) = \frac{\alpha\mathrm{e}^{-\Delta(1+\lambda)/c}}{c|q'(\Delta)|}\left(\frac{1-\mathrm{e}^{-\Delta(\alpha-1)/c}}{\alpha-1}\right). \tag{5.24}$$

Moreover, we have simply that $-cq'(\phi) = -h + \psi(\phi) - ga(\phi)$ for $\phi \in \{0, \Delta\}$. One natural way to find the zeros of $\mathcal{E}(\lambda)$ is to write $\lambda = \nu + i\omega$ and plot the zero contours of Re $\mathcal{E}(\lambda)$ and Im $\mathcal{E}(\lambda)$ in the (ν, ω) plane. The Evans function is zero where the lines intersect. We do precisely this in Figure 5.2 for three distinct points on the solution branch shown in Figure 5.1. On the fast branch it would appear that all the zeros of the Evans function lie in the left-hand complex plane, while for the slow wave there is at least one in the right-hand plane (on the real axis). As expected there is a double zero eigenvalue as one passes from the fast to the slow branch of traveling pulse solutions. Hence, the fast wave is stable and the slow wave unstable, confirming the numerical observations made in [13]. We note that in the presence of a discrete synaptic transmission delay of duration τ_d, we have the replacement $\eta(t) \to \eta(t-\tau_d)$. This causes the corresponding changes $A(\xi,\lambda) \to A(\xi + c\tau_d, \lambda)\mathrm{e}^{-\lambda\tau_d}$, etc., and it is a simple matter to recalculate expressions (5.20)–(5.24). Although discrete delays influence the speed of solution, a direct examination of the Evans function in this case shows that they do not induce any new instabilities.

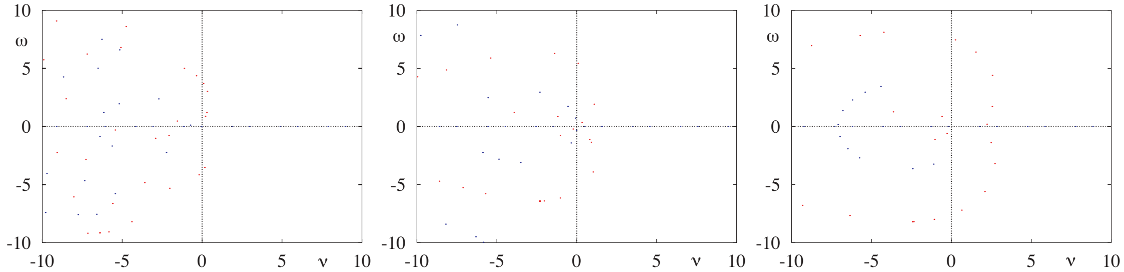


Figure 5.2. *Evans function for a traveling pulse in a model with nonlinear recovery. Red lines indicate where* Re $\mathcal{E}(\lambda) = 0$ *and blue where* Im $\mathcal{E}(\lambda) = 0$*. Zeros of the Evans function occur at the intersection of red and blue lines. In the left-hand figure,* $g = 2$ *and a solution is taken from the fast branch. In the middle, the value of* g *is that at the saddle-node bifurcation from Figure* 5.1*. On the right,* $g = 2$ *with a solution taken from the slow branch. Other parameters are* $h = 0.1$*,* $\alpha = 2$*, and* $v = 10$*.*

5.2. Example: A dynamic instability of a standing pulse. In this section we choose $\eta(t) = \alpha e^{-\alpha t}$, $\eta_a(t) = e^{-t}$, $w(x) = e^{-|x|}/2$, and $w_a(x) = e^{-|x|/\sigma_a}/(2\sigma_a)$ to recover a model of Pinto and Ermentrout [35]. In this case the standing pulse is given by

$$q(\xi) = \begin{cases} \frac{1}{2}(e^{\xi} - e^{\xi-\Delta}) - \frac{g}{2}(e^{\xi/\sigma_a} - e^{(\xi-\Delta)/\sigma_a}), & \xi \le 0, \\ 1 - g - \frac{1}{2}(e^{\xi-\Delta} + e^{-\xi}) + \frac{g}{2}(e^{(\xi-\Delta)/\sigma_a} + e^{-\xi/\sigma_a}), & 0 < \xi < \Delta, \\ \frac{1}{2}(e^{-(\xi-\Delta)} - e^{-\xi}) - \frac{g}{2}(e^{-(\xi-\Delta)/\sigma_a} - e^{-\xi/\sigma_a}), & \xi \ge \Delta. \end{cases} \tag{5.25}$$

Enforcing the condition $q(0) = h$ or $q(\Delta) = h$ generates the pulse width as a function of system parameters:

$$\frac{1}{2}(1 - e^{-\Delta}) - \frac{g}{2}(1 - e^{-\Delta/\sigma_a}) = h. \tag{5.26}$$

A plot of the pulse width as a function of the threshold parameter h is shown in Figure 5.3, showing that solutions come in pairs. We may then use (5.15) and (5.16) to construct the Evans function and plot it in the same fashion as the last example. However, unlike the last example we find that there is not a simple exchange of stability as one passes through the limit point defining the transition from a broad to a narrower pulse. Indeed we see from Figure 5.4 that it is possible for a solution on the upper branch of Figure 5.3 to undergo a *dynamic* instability with increasing α. By dynamic we mean that a pair of complex eigenvalues crosses into the right-hand plane on the imaginary axis so that the standing pulse may begin to oscillate, as originally described in [35].

For the parameter values in Figure 5.4 and choosing a value of α below that defining a dynamic instability, direct numerical simulations show that a bump solution is stable to random perturbations. In contrast, beyond the dynamic instability point, a bump solution can destabilize in favor of a homogeneous steady state. These two cases are illustrated in Figure 5.5. The critical value of α defining a dynamic instability is found to depend only weakly on the value of the axonal conduction velocity v.

For small values of the threshold h the bump solution can develop a *dimple* such that $q''(0) > 0$. We plot the Evans function for a dimple solution in Figure 5.6 and note that it also

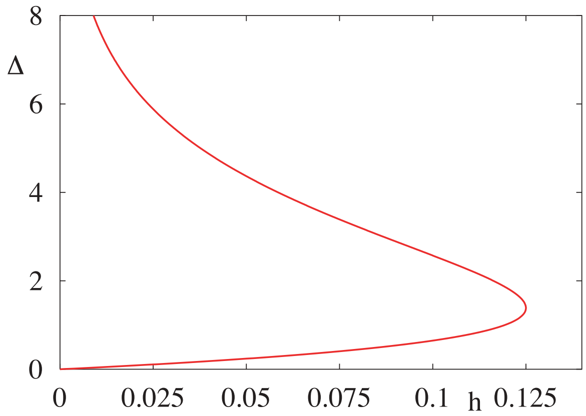


Figure 5.3. *Pulse width as a function of threshold h in a model with lateral inhibition and nonlinear recovery. Here $\sigma_a = 2$ and $g = 1$.*

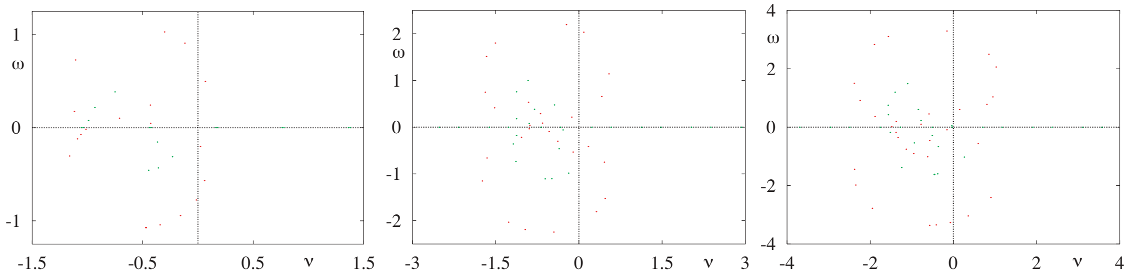


Figure 5.4. *Evans function for the model with lateral inhibition and nonlinear recovery. Here $h = 0.1$ and $v = 10$ and a solution is taken from the branch with largest width Δ. On the left $\alpha = 0.5$, and in the middle $\alpha = 1.0$, while on the right $\alpha = 1.5$. This illustrates the possibility of a dynamic instability with increasing α as a pair of complex eigenvalues crosses over to the right-hand plane through the imaginary axis.*

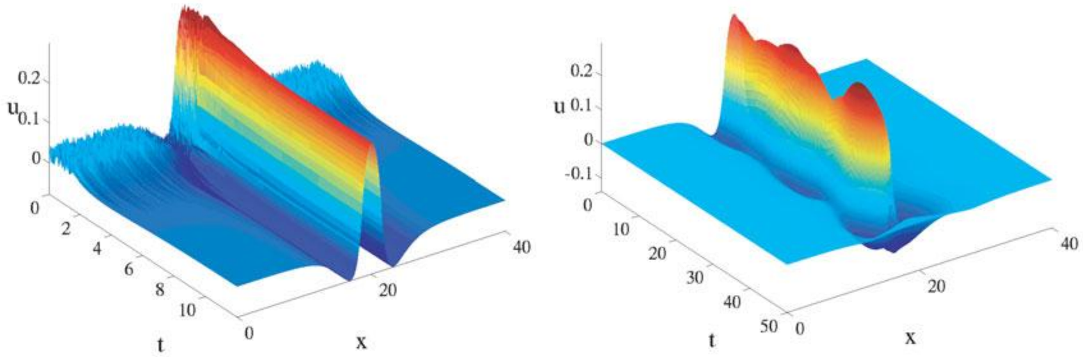


Figure 5.5. 3-d *plot of stable bump ($\alpha = 0.8$) and destabilized bump ($\alpha = 1.1$). $v = 1$, $h = 0.1$. The first has noisy initial data, with rapid convergence to the stable bump solution. The second case has initial data with $u(x, 0) = 1.05q(x)$, where $q(x)$ is the stationary bump solution. Clicking on the above images displays the associated movie* (60595_03.mov).

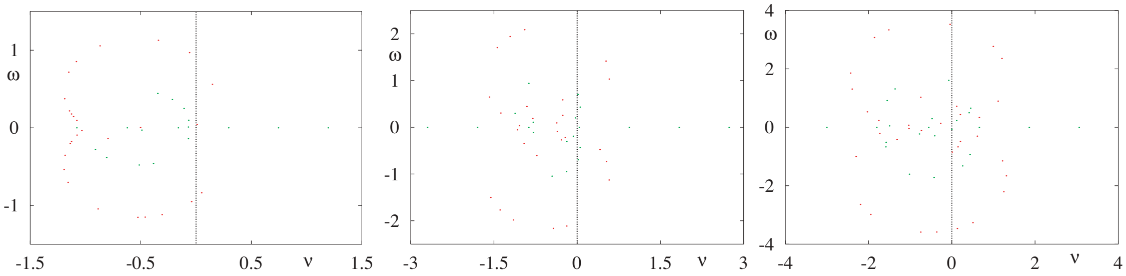


Figure 5.6. *Evans function for the model with lateral inhibition and nonlinear recovery. Here* $h = 0.025$ *and* $v = 1$ *and a solution is taken from the branch with largest width* Δ*. On the left* $\alpha = 0.5$*, and in the middle* $\alpha = 0.9$*, while on the right* $\alpha = 1.5$*. This illustrates the possibility of a dynamic instability with increasing* α *as a pair of complex eigenvalues crosses over to the right-hand plane through the imaginary axis.*

undergoes a dynamic instability with increasing α. Interestingly, direct numerical simulations show that in this case the value of v can have a more profound effect on the dynamics beyond the point of instability. For large values of v an unstable solution collapses to a homogeneous steady state, whereas lower values of v lead to the shedding of a pair of left and right traveling pulses. This is illustrated in Figure 5.7.

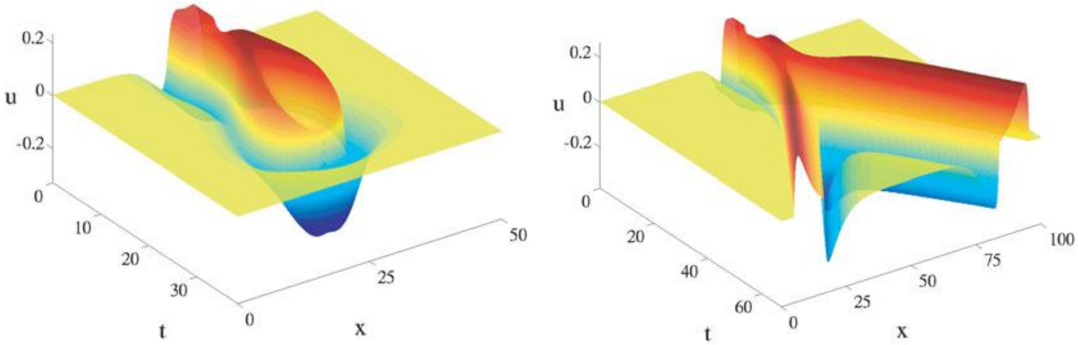


Figure 5.7. $h = 0.025, \alpha = 1.1,$ *and the bump width is* 5.88*. Left:* $v = 8$*; the bump destabilizes and dies. Right:* $v = 1$*; as the bump dies, it sheds a pair of pulse waves. Clicking on the above images displays the associated movies* (60595_04.mov *and* 60595_05.mov)

6. Discussion. In this paper we have shown how to calculate the Evans function for integral neural field equations with a Heaviside firing rate function. Our work generalizes that of Zhang [41] on a certain integro-differential equation model in a number of ways. These include (i) studying exponential wave stability in an integral framework rather than an integro-differential framework, (ii) avoiding the need to resort to the study of some singularly perturbed system, and (iii) including the effects of space-dependent delays arising from axonal communication. For the three main models that we have considered, i.e., a scalar integral neural field, a model with linear recovery, and a model with nonlinear recovery, we have presented the explicit form for the Evans function for traveling fronts and pulses. Moreover, through a variety of examples we have shown that this is a powerful tool for the stability analysis of nonlinear waves and localized patterns in integral neural field models. In all cases the Evans function $\mathcal{E}(\lambda)$ is a complex analytic function of $\lambda \in \mathbb{C}$, with all the usual properties expected of such a function (although we have not presented a *rigorous* proof of that here). Namely, $\mathcal{E}(\lambda) = 0$ if and only if λ is an eigenvalue, the order of the roots is equal to the multiplicity of eigenvalues, and for Re $\lambda > 0$, $\lim_{|\lambda|\to\infty} \mathcal{E}(\lambda) = 1$. This means that there is a positive constant M such that $\mathcal{E}(\lambda) \neq 0$ for all $|\lambda| \geq M$. Since $\mathcal{E}(\lambda)$ is complex analytic, there are at most finitely many eigenvalues within the disc $|\lambda| = M$. One could, of course, resort to the computation of $\mathcal{E}'(\lambda)/\mathcal{E}(\lambda)$ along the imaginary axis and use the argument principle to determine the number of zeros in the right half plane. Alternatively we could construct a Nyquist plot (the image of the Evans function along the imaginary axis) and count the number of times the graph winds around the origin. However, in this paper we have chosen simply to graphically find the intersection of the zero contours of Re $\mathcal{E}(\lambda)$ and Im $\mathcal{E}(\lambda)$ in the complex plane.

There are a number of natural ways in which to extend the work presented in this paper. Perhaps the most natural extension is to consider the issue of a forced neural field where waves may lock to a moving stimulus. Another is to consider the use of averaging and homogenization theory to uncover the role of the periodic microstructure of cortex in front and pulse propagation and its failure, along the lines developed in [5, 7]. This is especially important when one recalls that the traveling front and pulse solutions considered in this paper are not structurally stable so that the introduction of even small inhomogeneities in the connectivity pattern may lead to propagation failure. Apart from recent work by Taylor [37], Werner and Richter [38], Laing and Troy [29], and Folias and Bressloff [18], planar studies have also received relatively little attention. Many of the techniques in this paper will carry over to the case of radially symmetric solutions, although the study of, say, spiral stability would first require the explicit construction of such solutions. However, the most obvious and major challenge is to extend this work to cover smooth sigmoidal firing rate functions. Even the numerical construction of the Evans function in this case is likely to be highly nontrivial, although there is some hope that recent Magnus methods developed by Aparicio, Malham, and Oliver may be suited to this task [3]. These and other topics are ongoing areas of current research and will be reported on elsewhere.

Acknowledgments. The authors would like to thank L. Zhang for making available early drafts of papers on the stability of traveling wave solutions in synaptically coupled neuronal networks. S. Coombes would like to acknowledge the invitation from David Terman and Björn Sandstede to attend the workshop *Nonlinear Integro-Differential Equations in Mathematics and Biology*, March 3–5, 2003, at the Mathematical Biosciences Institute, Ohio State University, which directly motivated the work in this paper.

REFERENCES

[1] J. Alexander, R. Gardner, and C. Jones, *A topological invariant arising in the stability analysis of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] S. Amari, *Dynamics of pattern formation in lateral-inhibition type neural fields*, Bio. Cybernet., 27 (1977), pp. 77–87.

[3] N. D. Aparicio, S. J. A. Malham, and M. Oliver, *Numerical evaluation of the Evans function by Magnus integration*, BIT, submitted.

[4] N. J. Balmforth, R. V. Craster, and S. J. A. Malham, *Unsteady fronts in an autocatalytic system*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 455 (1999), pp. 1401–1433.

[5] P. C. Bressloff, *Traveling fronts and wave propagation failure in an inhomogeneous neural network*, Phys. D, 155 (2001), pp. 83–100.

[6] P. C. Bressloff and S. E. Folias, *Front bifurcations in an excitatory neural network*, SIAM J. Appl. Math., 65 (2004), pp. 131–151.

[7] P. C. Bressloff, S. E. Folias, A. Prat, and Y. X. Li, *Oscillatory waves in inhomogeneous neural media*, Phys. Rev. Lett., 91 (2003), pp. 178101-1–178101-4.

[8] T. J. Bridges, G. Derks, and G. Gottwald, *Stability and instability of solitary waves of the fifth-order KdV equation: A numerical framework*, Phys. D, 172 (2002), pp. 190–216.

[9] F. Chen, *Travelling waves for a neural network*, Electron. J. Differential Equations, 2003 (2003), pp. 1–4.

[10] X. Chen, *Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution equations*, Adv. Differential Equations, 2 (1997), pp. 125–160.

[11] Z. Chen, G. B. Ermentrout, and J. B. McLeod, *Traveling fronts for a class of nonlocal convolution differential equation*, Appl. Anal., 64 (1997), pp. 235–253.

[12] R. D. Chervin, P. A. Pierce, and B. W. Connors, *Propagation of excitation in neural network models*, J. Neurophysiology, 60 (1988), pp. 1695–1713.

[13] S. Coombes, G. J. Lord, and M. R. Owen, *Waves and bumps in neuronal networks with axo-dendritic synaptic interactions*, Phys. D, 178 (2003), pp. 219–241.

[14] D. Cremers and A. V. M. Herz, *Traveling waves of excitation in neural field models: Equivalence of rate descriptions and integrate-and-fire dynamics*, Neural Computation, 14 (2002), pp. 1651–1667.

[15] G. B. Ermentrout, *Neural nets as spatio-temporal pattern forming systems*, Rep. Progr. Phys., 61 (1998), pp. 353–430.

[16] G. B. Ermentrout and J. B. McLeod, *Existence and uniqueness of travelling waves for a neural network*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 461–478.

[17] J. Evans, *Nerve axon equations* IV: *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.

[18] S. E. Folias and P. C. Bressloff, *Breathing pulses in an excitatory neural network*, SIAM J. Appl. Dynam. Sys., 3 (2004), pp. 378–407.

[19] D. Golomb and Y. Amitai, *Propagating neuronal discharges in neocortical slices: Computational and experimental study*, J. Neurophysiology, 78 (1997), pp. 1199–1211.

[20] A. Hagberg and E. Meron, *Pattern formation in non-gradient reaction-diffusion systems: The effects of front bifurcations*, Nonlinearity, 7 (1994), pp. 805–835.

[21] P. Howard and K. Zumbrun, *The Evans function and stability criteria for degenerate viscous shock waves*, Discrete Contin. Dyn. Syst. Ser. A, 4 (2004), pp. 837–855.

[22] A. Hutt, M. Bestehorn, and T. Wennekers, *Pattern formation in intracortical neuronal fields*, Network, 14 (2003), pp. 351–368.

[23] M. A. P. Idiart and L. F. Abbott, *Propagation of excitation in neural network models*, Network, 4 (1993), pp. 285–294.

[24] C. K. R. T. Jones, *Stability of the traveling wave solutions of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.

[25] T. Kapitula, N. Kutz, and B. Sandstede, *The Evans function for nonlocal equations*, Indiana Univ. Math. J., to appear.

[26] T. Kapitula and B. Sandstede, *Edge bifurcations for near integrable systems via Evans function techniques*, SIAM J. Math. Anal., 33 (2002), pp. 1117–1143.

[27] T. Kapitula and B. Sandstede, *Eigenvalues and resonances using the Evans function*, Discrete Contin. Dynam. Systems, 10 (2004), pp. 857–869.

[28] U. Kim, T. Bal, and D. A. McCormick, *Spindle waves are propagating synchronized oscillations in the ferret LGNd in vitro*, J. Neurophysiology, 74 (1995), pp. 1301–1323.

[29] C. R. Laing and W. C. Troy, *PDE methods for nonlocal models*, SIAM J. Appl. Dynam. Sys., 2 (2003), pp. 487–516.

[30] Y. A. Li and K. Promislow, *The mechanism of the polarization mode instability in birefringent fiber optics*, SIAM J. Math. Anal., 31 (2000), pp. 1351–1373.

[31] R. Miles, R. D. Traub, and R. K. S. Wong, *Spread of synchronous firing in longitudinal slices from the CA*3 *region of the hippocampus*, J. Neurophysiology, 60 (1988), pp. 1481–1496.

[32] R. L. Pego and M. I. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.

[33] R. L. Pego and M. I. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.

[34] 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.

[35] 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.

[36] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems, Vol. 2, North–Holland, Amsterdam, 2002, pp. 983–1055.

[37] J. G. Taylor, *Neural 'bubble' dynamics in two dimensions: Foundations*, Biol. Cybernet., 80 (1999), pp. 393–409.

[38] H. Werner and T. Richter, *Circular stationary solutions in two-dimensional neural fields*, Biol. Cybernet., 85 (2001), pp. 211–217.

[39] H. R. Wilson and J. D. Cowan, *A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue*, Kybernetik, 13 (1973), pp. 55–80.

[40] J.-Y. Wu, L. Guan, and Y. Tsau, *Propagating activation during oscillations and evoked responses in neocortical slices*, J. Neuroscience, 19 (1999), pp. 5005–5015.

[41] L. Zhang, *On stability of traveling wave solutions in synaptically coupled neuronal networks*, Differential Integral Equations, 16 (2003), pp. 513–536.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 601–619

Disease Induced Oscillations between Two Competing Species*

P. van den Driessche† and M. L. Zeeman‡

Abstract. The interaction of disease and competition dynamics is investigated in a system of two competing species in which only one species is susceptible to disease. The model is kept as simple as possible, combining Lotka–Volterra competition between the species with disease dynamics of susceptible and infective individuals within one of the species. It is assumed that pure vertical disease transmission (from parent to offspring) dominates horizontal transmission (by contact between infective and susceptible individuals) and that infective individuals have the same competition strength as susceptibles but a lower intrinsic growth rate. These assumptions yield three-dimensional competitive Lotka–Volterra dynamics modeling the disease-competition interaction. It is proved that if in the absence of disease there is competitive exclusion between the two species, then the presence of disease can lead to stable or oscillatory coexistence of both species. The case of oscillatory coexistence can be viewed either as disease induced oscillations between competing species or as competition induced oscillations in an endemic disease. By contrast, conditions are found under which, if the two species coexist in the absence of disease, then the introduction of disease does not induce oscillations, and the long-term dynamics are determined by the basic reproduction number.

Key words. Lotka–Volterra competition, SI disease, Hopf bifurcation, periodic orbit, basic reproduction number

AMS subject classifications. 92D25, 92D30, 37C27

DOI. 10.1137/030600394

1. Introduction.

We investigate how disease and competition dynamics interact in a system of two competing species, in which only species 1 is susceptible to disease. We keep the model as simple as possible, combining disease within species 1 with Lotka–Volterra competition between the species. This yields a three-dimensional competitive Lotka–Volterra system modeling the disease-competition interaction. We prove that if species 1 can drive species 2 to extinction in the absence of disease, then the introduction of disease can weaken species 1 sufficiently to permit stable or oscillatory coexistence of both species.

The case of oscillatory coexistence can be viewed in two ways: as disease induced oscillations in competing populations or as competition induced oscillations in an endemic disease. Oscillations are observed in the incidence data of several diseases. Factors such as periodic coefficients, nonlinear incidence, time delays, and stochasticity can give rise to oscillations in diseases models; see, for example, Hethcote and Levin [10]. However, in the model analyzed here, oscillations arise purely as a result of interaction between competition and disease, without the inclusion of other factors.

*Received by the editors June 30, 2003; accepted for publication (in revised form) December 23, 2003; published electronically December 27, 2004.
http://www.siam.org/journals/siads/3-4/60039.html

†Department of Mathematics and Statistics, University of Victoria, Victoria V8W 3P4, BC, Canada (pvdd@math.uvic.ca). This research was supported in part by an NSERC Discovery Grant.

‡Department of Applied Mathematics, University of Texas at San Antonio, San Antonio, TX 78249 (zeeman@math.utsa.edu). This research was supported in part by NSF IGERT grant DGE-9870631 at Cornell University and NSF IGMS grant DMS-0308956.

Our emphasis is on proving global results about the qualitative behavior of the model. Most similar models considered in the literature rely on linear analysis coupled with numerical simulation. In a seminal paper on invasions by infectious diseases, Anderson and May [1] considered a similar model (equation (26), p. 557, and Table 5) in which a pathogen invades and infects only one competing host. They predicted, without proof, disease induced stable coexistence of the competing species but did not predict any oscillatory behavior. They also cited several ecological examples of disease-competition interactions—for example, native and introduced bird species in Hawaii, native and introduced squirrels in Britain, the influence of myxoma virus on hares and rabbits in Europe, and a sporozoan parasite in competing species of flour beetle. In all cases, the disease is significantly more pathogenic in one of the species than the other. More recently Venturino [22] considered another similar three-dimensional model, proved that locally stable disease induced coexistence is possible, and presented numerical evidence of disease induced oscillations in populations. Some host-host-pathogen models were summarized by Begon and Bowers [2]. One early such model was by Holt and Pickering [14], but they assumed no direct competition either within or between species. The introduction of density dependence in the Holt–Pickering model was found by Greenman and Hudson [8] to lead to a variety of outcomes, including oscillatory behavior. Both competitive and infective interactions for a directly transmitted disease were considered by Bowers and Turner [3, equations (1)–(4)]. Using feasibility and linear stability arguments, they discussed conditions for infected coexistence. A two-host susceptible-infective (SI) model with competition was considered by Greenman and Hudson [9]. They used a geometric approach to construct bifurcation maps and found the possibility of both coexistence and single-host equilibria (the outcome being initial-condition–dependent).

To introduce our model, we recall two-dimensional competitive Lotka–Volterra system dynamics (section 2) and then include a disease with susceptible and infective individuals within species 1, leading to a three-dimensional competitive Lotka–Volterra model (section 3, system (3) with inequalities (4)). In section 4, we introduce a basic reproduction number $\mathcal{R}_{01}$ for disease in species 1 in the absence of species 2. Using a geometric approach based on nullcline analysis and the monotonicity results of Hirsch [11], three-dimensional competitive Lotka–Volterra systems have been classified by Zeeman [24] into 33 equivalence classes (see Figure 3 in section 5). Geometric ideas from [11, 18] and [24] are central to our analysis and are summarized for our model in section 5.

In sections 6 and 7, we prove that if in the absence of disease species 1 drives species 2 to extinction, or there is initial-condition–dependent competitive exclusion between the species, then the introduction of disease can weaken species 1 sufficiently to give rise to stable or oscillatory coexistence of the two species. By contrast, in section 8 we find an inequality (20) under which, if the two species coexist in the absence of disease, then the introduction of disease does not induce oscillations. In sections 6 and 7, the persistence of endemic disease in the full model (either stable or oscillatory) depends only on the value of the basic reproduction number, $\mathcal{R}_{01}$, for the disease in species 1 in the absence of species 2. In section 8, where the two species coexist in the absence of disease, the basic reproduction number for the full system, $\mathcal{R}_0$, is needed to characterize persistence of the disease. We show that if $\mathcal{R}_0 < 1$, then the equilibrium representing disease-free coexistence of the species is globally asymptotically stable. When $\mathcal{R}_0 > 1$ and inequality (20) is satisfied, there is a globally asymptotically stable

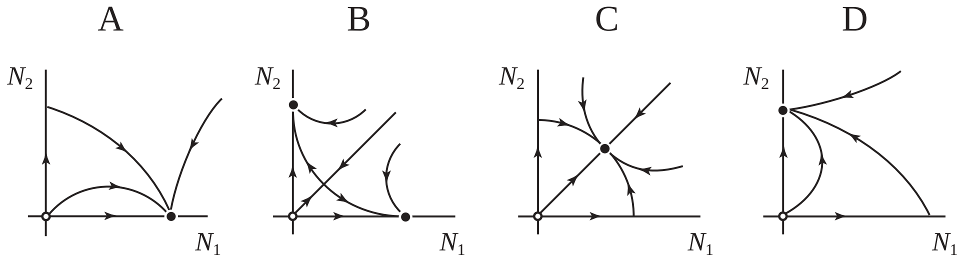


Figure 1. *Qualitative dynamics of Cases* A–D*. An attracting equilibrium is denoted by a closed bullet, a repelling equilibrium by an open bullet, and a saddle equilibrium by the intersection of its stable and unstable manifolds.*

equilibrium representing coexistence of the species, with endemic disease in species 1. In section 9 we prove that, as expected, the introduction of disease does not change the long-term demographics if species 2 drives species 1 to extinction in the absence of disease.

2. Disease-free competition. We assume that in the absence of disease, the competition between the two species is modeled by the competitive Lotka–Volterra system:

$$\begin{aligned} N_1' &= N_1(r_1 - a_{11}N_1 - a_{12}N_2), \\ N_2' &= N_2(r_2 - a_{21}N_1 - a_{22}N_2). \end{aligned} \tag{1}$$

Here N_i denotes the population size of species i, and the prime denotes differentiation with respect to time. For each i, j, the constant r_i represents the combined intrinsic birth and death rates of species N_i, and the coefficient a_{ij} represents the competitive impact of species j on the growth of species i. We assume throughout that r_i and a_{ij} are strictly positive. We define the ith *nullcline* of system (1) to be the line $\sum_{j=1}^{2} a_{ij}N_j = r_i$ on which $N_i' = 0$. The determinant of the coefficient matrix $\left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$ is denoted by

$$\Delta = a_{11}a_{22} - a_{12}a_{21}. \tag{2}$$

The long-term dynamics of system (1) are well understood; see, for example, [12, section 3.3]. There are four qualitatively different, nondegenerate phase portraits, denoted throughout as Cases A–D, as shown in Figure 1. Table 1 characterizes each case by the global dynamics, by the geometric configuration of the nullclines, and by the algebraic inequalities among the parameters. One way to translate between the algebraic inequalities and the nullcline configuration is to interpret each inequality as an ordering of the intersections of the two nullclines with a particular coordinate axis. See [12, section 3.3], [19, section 3.5], or [24] for more detail. It is easy to see from the algebraic inequalities that Δ is negative in Case B, positive in Case C, and can take any sign in Cases A and D. In sections 6–9 we discuss the consequences of introducing disease to species 1 for each of Cases A–D, respectively. The examples given in Table 1 are used to develop Examples A–C in sections 6–8.

3. Competition with disease in species 1. We model disease in species 1 by dividing the population N_1 into two compartments: susceptibles S and infectives I. Figure 2 illustrates

Table 1
Table of Cases A–D.

Case	Nullclines in $\mathbf{R}^2_+$	Inequalities	Example
A. N_1 drives N_2 to extinction	Nullclines disjoint: N_1 entirely above N_2	$r_2a_{11} < r_1a_{21}$ $r_2a_{12} < r_1a_{22}$	$N_1' = N_1(25 - 10N_1 - 10N_2)$ $N_2' = N_2(20 - 9N_1 - 8.5N_2)$
B. Initial cond. dependent comp. exclusion	Nullclines intersect: N_1 above N_2 on N_1-axis N_2 above N_1 on N_2-axis	$r_2a_{11} < r_1a_{21}$ $r_1a_{22} < r_2a_{12}$	$N_1' = N_1(25 - 10N_1 - 17N_2)$ $N_2' = N_2(20 - 8.5N_1 - 12N_2)$
C. Stable coexistence	Nullclines intersect: N_2 above N_1 on N_1-axis N_1 above N_2 on N_2-axis	$r_1a_{21} < r_2a_{11}$ $r_2a_{12} < r_1a_{22}$	$N_1' = N_1(20 - 15N_1 - 10N_2)$ $N_2' = N_2(15 - 10N_1 - 20N_2)$
D. N_2 drives N_1 to extinction	Nullclines disjoint: N_2 entirely above N_1	$r_1a_{21} < r_2a_{11}$ $r_1a_{22} < r_2a_{12}$	$N_1' = N_1(20 - 9N_1 - 8.5N_2)$ $N_2' = N_2(25 - 10N_1 - 10N_2)$

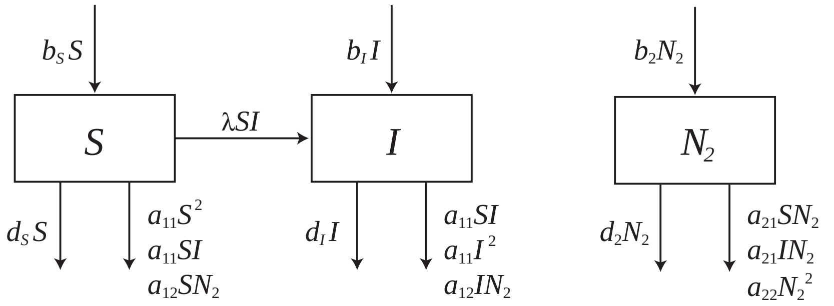


Figure 2. *Compartmental diagram of disease-competition interaction.*

the mechanisms for growth and decline of the S and I populations over many generations, under the influence of the disease and competition with species 2.

We assume there is horizontal transmission of the disease with simple mass action incidence with λ as the mass action coefficient. This incidence assumes that the contact rate increases linearly with each population size; see, for example, [10, p. 201]. Also we assume that there is pure vertical transmission, in the sense that all births to infective individuals are infective; see [4] for a comprehensive description of models that include vertical transmission. The susceptible and infective populations have intrinsic birth rates b_S and b_I and intrinsic death rates d_S and d_I, respectively. The intrinsic birth and death rates of susceptible individuals into and out of the susceptible class are assumed to be the same as for the disease-free situation, so $r_S = b_S - d_S = r_1$. By contrast, we assume there may be disease induced death ($d_S < d_I$), or disease induced infertility ($b_I < b_S$), but not enough for deaths to outweigh births. So $0 < b_I - d_I = r_I < r_S$. In fact, we strengthen this assumption to $(a_{11} + \lambda)r_I < a_{11}r_S$, so that the unrealistic case of 100% disease is ruled out; this is further explained in section 4. Increasing either of the disease parameters λ or r_I has the effect of strengthening the disease, thereby weakening species 1.

Combining the disease and competition models, we assume that the disease does not weaken infective individuals for competition. Thus the competitive impact of the susceptible and infective populations on the growth rate of each other has coefficient a_{11}, while their competitive impact on species 2 has coefficient a_{21}. Similarly, the competitive impact of species 2 on each of the susceptible and infective populations has coefficient a_{12}. Finally, horizontal disease transmission is assumed to have a weaker impact than intraspecific competition so that $\lambda < a_{11}$.

The assumptions listed above yield the ordinary differential equations

$$\begin{aligned} S' &= S(r_S - a_{11}S - (a_{11}+\lambda)I - a_{12}N_2), \\ I' &= I(r_I - (a_{11}-\lambda)S - a_{11}I - a_{12}N_2), \\ N_2' &= N_2(r_2 - a_{21}S - a_{21}I - a_{22}N_2) \end{aligned} \tag{3}$$

and ensure that

$$0 < (a_{11}+\lambda)r_I < a_{11}r_S, \quad 0 < r_2, \quad 0 < \lambda < a_{11}, \text{ and } 0 < a_{ij} \text{ for each } i,j = 1,2. \tag{4}$$

Hence system (3) is a three-dimensional competitive Lotka–Volterra system. In particular, the positive orthant is invariant, and there is a compact attracting region. Solutions having nonnegative initial values remain nonnegative for all further time, are asymptotic to solutions on the carrying simplex (see section 5), and thus eventually satisfy $0 \le S \le r_S/a_{11}$, $0 \le I \le r_I/a_{11}$, $0 \le N_2 \le r_2/a_{22}$. Note that in the absence of disease (setting $I = 0$, $S = N_1$), system (3) reduces to system (1) since $r_S = r_1$.

4. SI dynamics in the absence of species 2. In the absence of species 2, system (3) with inequalities (4) reduces to an SI disease model with horizontal and pure vertical transmission that includes density-dependent death,

$$\begin{aligned} S' &= S(r_S - a_{11}S - (a_{11}+\lambda)I), \\ I' &= I(r_I - (a_{11}-\lambda)S - a_{11}I), \end{aligned} \tag{5}$$

where

$$0 < (a_{11}+\lambda)r_I < a_{11}r_S \text{ and } 0 < \lambda < a_{11}. \tag{6}$$

System (5) with inequalities (6) is also a two-dimensional competitive Lotka–Volterra system, so the global dynamics are easily understood, as in section 2. It is helpful for us to reinterpret those dynamics here, from the point of view of disease.

The disease-free equilibrium (DFE) for system (5) is given by $S = r_S/a_{11}$, $I = 0$. The basic reproduction number for the disease in species 1, determined from the next generation matrix of system (5) at the DFE (see, for example, [21]), is defined by

$$\mathcal{R}_{01} = \frac{\lambda S + b_I}{a_{11}S + d_I} = \frac{\lambda(r_S/a_{11}) + b_I}{r_S + d_I}. \tag{7}$$

Here, $\mathcal{R}_{01}$ is the average number of new infections (from horizontal and vertical transmission) caused by one infective during its average infectious period when introduced into a fully susceptible population of species 1. The term involving λ gives the contribution from the horizontal transmission, whereas the term involving b_I comes from vertical transmission, and $1/(r_S + d_I)$ is the average time in I when the population is fully susceptible. Models with vertical transmission and density-dependent demographics are considered by Busenberg and

Cooke [4, section 2.11] and Gao and Hethcote [7, section 7], where they derive nondimensional parameters similar to $\mathcal{R}_{01}$.

Note that since $r_I = b_I - d_I$,

$$\mathcal{R}_{01} < 1 \Leftrightarrow a_{11} r_I < r_S(a_{11} - \lambda). \tag{8}$$

Lemma 4.1. *For system* (5) *with inequalities* (6)*, if* $\mathcal{R}_{01} < 1$*, then the DFE on the* S *axis, namely,* $(S, I) = (r_S/a_{11}, 0)$*, is globally asymptotically stable in* $\text{int}\mathbf{R}^2_+$*.*

Proof. The I and S nullclines intersect the S axis at $r_I/(a_{11} - \lambda)$ and r_S/a_{11}, respectively. By (8)

$$\mathcal{R}_{01} < 1 \Rightarrow \frac{r_I}{a_{11} - \lambda} < \frac{r_S}{a_{11}},$$

so the S nullcline meets the S axis above the I nullcline. The I and S nullclines intersect the I axis at r_I/a_{11} and $r_S/(a_{11} + \lambda)$, respectively. So, by the assumption that $(a_{11}+\lambda)r_I < a_{11}r_S$ from (6), the S nullcline meets the I axis above the I nullcline. Thus the S nullcline lies entirely above the I nullcline in $\mathbf{R}^2_+$. So there is no equilibrium in $\text{int}\mathbf{R}^2_+$, and the DFE on the S axis is globally asymptotically stable in $\text{int}\mathbf{R}^2_+$ (see Table 1). ■

Lemma 4.2. *For system* (5) *with inequalities* (6)*, if* $\mathcal{R}_{01} > 1$*, then the endemic equilibrium, namely,* $(S, I) = (1/\lambda^2)(r_S a_{11} - r_I(a_{11} + \lambda),\ r_I a_{11} - r_S(a_{11} - \lambda))$*, is globally asymptotically stable in* $\text{int}\mathbf{R}^2_+$*.*

Proof. The assumption that $(a_{11} + \lambda)r_I < a_{11}r_S$ ensures that the I nullcline intersects the I axis below the S nullcline, so the equilibrium on the I axis is not locally attracting [24, p. 199], thereby ruling out the possibility of 100% infectives. Thus, when $\mathcal{R}_{01} > 1$ in system (5) with inequalities (6), there is a globally attracting equilibrium in the interior of the positive quadrant, representing stable endemic disease in species 1. ■

Note that decreasing either of the disease parameters λ or $r_I = b_I - d_I$ decreases $\mathcal{R}_{01}$. Indeed, a straightforward computation shows that either of the disease parameters can be used to control the disease by reducing $\mathcal{R}_{01}$ to below 1.

5. Geometric preliminaries.

The carrying simplex. The origin is a repelling equilibrium of system (3). Following [11] and [24], we define the *carrying simplex*, Σ, to be the boundary in $\mathbf{R}^3_+$ of the basin of repulsion of the origin. Exploiting the backward time monotonicity of system (3), a theorem of Hirsch [11, Theorem 1.7] shows that the carrying simplex is a globally attracting set, in the sense that every trajectory in $\mathbf{R}^3_+ \setminus 0$ is asymptotic to one in Σ, and that Σ is geometrically and topologically simple, in the sense that it is a two-dimensional Lipschitz submanifold homeomorphic to the standard unit simplex $(S + I + N_2 = 1)$ in $\mathbf{R}^3_+$ by radial projection. Thus the long-term behavior of system (3) is determined by the dynamics on Σ, and the nonzero forward limit sets in $\mathbf{R}^3_+$ all lie on Σ. The Poincaré–Bendixson theorem holds for the restriction of the system to Σ, so the only forward limit sets are equilibria and cycles. See [11, 24] for more details.

In Figures 4–7, the dynamics on the carrying simplex, Σ, are visualized for a variety of numerical examples of system (3). These figures were generated using the programs *CSimplex* [15] and *Geomview* [16]. The carrying simplex is viewed from above—in other words, from a point in the positive cone with coordinates much larger than any of the coordinates of points in Σ. In Figure 3, the dynamics on the carrying simplex are viewed from above, after radial projection to the standard unit simplex in $\mathbf{R}^3_+$.

Equilibria. It is easy to see that there are three axial equilibria at the vertices of Σ:

$$V_S = (r_S/a_{11}, 0, 0), \quad V_I = (0, r_I/a_{11}, 0), \quad \text{and} \quad V_{N_2} = (0, 0, r_2/a_{22}).$$

There are also three equilibria in the coordinate planes:

$$\begin{aligned} Q_{SI} &= \frac{1}{\lambda^2}(r_S a_{11} - r_I(a_{11}+\lambda),\ r_I a_{11} - r_S(a_{11}-\lambda),\ 0),\\ Q_{SN_2} &= \frac{1}{\Delta}(r_S a_{22} - r_2 a_{12},\ 0,\ r_2 a_{11} - r_S a_{21}),\\ Q_{IN_2} &= \frac{1}{\Delta}(0,\ r_I a_{22} - r_2 a_{12},\ r_2 a_{11} - r_I a_{21}), \end{aligned}$$

where $\Delta = a_{11}a_{22} - a_{12}a_{21}$ as in (2). We say that Q_{SI} is biologically feasible if the S and I components are strictly positive. In that case, Q_{SI} lies on an edge of Σ and represents an endemic equilibrium for species 1. Recall, from section 4, that Q_{SI} is biologically feasible and globally attracting in the SI plane if and only if $\mathcal{R}_{01} > 1$. Similarly, when they are biologically feasible, Q_{SN_2} represents disease-free coexistence between the species, and Q_{IN_2} represents species coexistence with all species 1 being infective.

Let A be the coefficient matrix

$$A = \begin{pmatrix} a_{11} & a_{11}+\lambda & a_{12}\\ a_{11}-\lambda & a_{11} & a_{12}\\ a_{21} & a_{21} & a_{22} \end{pmatrix}.$$

$\mathrm{Det}(A) = \lambda^2 a_{22} > 0$, and so system (3) has an equilibrium at $P = (S^*, I^*, N_2^*) = A^{-1}(r_S, r_I, r_2)^T$. A straightforward computation shows that

$$\begin{aligned} S^* &= \frac{1}{a_{22}\lambda^2}(\Delta(r_S - r_I) - \lambda(a_{22}r_I - a_{12}r_2)),\\ I^* &= \frac{1}{a_{22}\lambda^2}(\lambda(a_{22}r_S - a_{12}r_2) - \Delta(r_S - r_I)),\\ N_2^* &= \frac{1}{a_{22}\lambda}(\lambda r_2 - a_{21}(r_S - r_I)), \end{aligned} \tag{9}$$

and $S^* + I^* = (r_S - r_I)/\lambda$. If $S^*, I^*, N_2^* > 0$, then $P \in \mathrm{int}\Sigma \subset \mathrm{int}\mathbf{R}^3_+$, and we say P is a biologically feasible coexisting endemic equilibrium. Note that by $(a_{11}+\lambda)r_I < a_{11}r_S$ of inequalities (4), if $N_2^* > 0$, then $S^* > 0$.

Nullclines. As in section 2, we define the S *nullcline* of system (3) to be the plane $r_S = a_{11}S + (a_{11}+\lambda)I + a_{12}N_2$, on which $S' = 0$. The I and N_2 nullclines are defined similarly. The following lemma, from [18], shows that for any pair of species to coexist, their nullclines must intersect in $\mathbf{R}^3_+$. If, for example, the S nullcline is disjoint from the I nullcline in $\mathbf{R}^3_+$ and lies in the unbounded component of $\mathbf{R}^3_+ \setminus \{I \text{ nullcline}\}$, then we say the S nullcline lies *entirely above* the I nullcline.

Lemma 5.1 (see [18, Lemma 3.1]). *For an n-dimensional competitive Lotka–Volterra system*

$$x_i' = x_i\left(r_i - \sum_{j=1}^{n} a_{ij}x_j\right), \quad \textit{with } r_i, a_{ij} > 0, \quad i = 1, \ldots, n,$$

if the x_i nullcline lies entirely above the x_j nullcline in $\mathbf{R}_+^n$, then x_i drives x_j to extinction.

Nullcline classes and periodic orbits. Figure 3 shows the 33 nullcline equivalence classes of three-dimensional competitive Lotka–Volterra systems defined by Zeeman [24]. The equivalence classes are defined modulo permutation of the populations (S, I, and N_2 in this context). Figure 3 has been redrawn from [24] so that the vertices can be interpreted as S: bottom left, I: bottom right, N_2: top. There are no periodic orbits in classes 1–25, 32, or 33, so the global dynamics are as shown [20, 24]. In class 33, there is stable coexistence of all the populations, while in classes 1–25 and 32, at least one population is driven to extinction. Exactly six classes, namely, 26–31, admit Hopf bifurcation to robust periodic orbits representing population oscillations [6, 20, 24]. When there are no periodic orbits, the global dynamics are as shown, except that the interior equilibrium in classes 26 and 27 may be either attracting or repelling [24]. Classes 26–29 admit at least two isolated periodic orbits [13, 17], but an upper bound on the number of isolated periodic orbits remains open [23]. We do not address the question of how many isolated periodic orbits can occur in system (3) under assumptions (4). However, in our simulations, no more than one periodic orbit has been observed. Not all of the nullcline classes are realized by system (3) under assumptions (4). In sections 6–8, we see examples from (at least) each of the classes 1, 5, 9, 12, 13, 14, 26, 29, and 33.

Using these geometric preliminaries, we now proceed to consider the introduction of disease to species 1 in each of Cases A–D of Figure 1 and Table 1. We numerically illustrate and interpret the dynamical behavior and prove that many of the qualitative features hold in general.

6. Case A.

For this section, assume that, in the absence of disease, species 1 drives species 2 to extinction, as in Case A of Figure 1 and Table 1. Introduction of disease is then modeled by system (3) with inequalities (4) and

$$r_2 a_{11} < r_S a_{21} \quad \text{and} \quad r_2 a_{12} < r_S a_{22} \tag{10}$$

from Table 1, Case A. We begin by describing a numerical example to illustrate the impact of disease in species 1 on the competition between the species.

Example A. Consider the system

$$\begin{aligned} S' &= S(25 - 10S - 15I - 10N_2), \\ I' &= I(r_I - 5S - 10I - 10N_2), \\ N_2' &= \tau N_2(20 - 9S - 9I - 8.5N_2), \ \tau > 0. \end{aligned} \tag{11}$$

The parameter r_I is used to change the strength of the disease, while τ is used to change the time scale of species 2 relative to species 1. The restriction of system (11) to the S, N_2 plane has a globally attracting fixed point at V_S (Figure 1, Case A), and the disease dynamics in

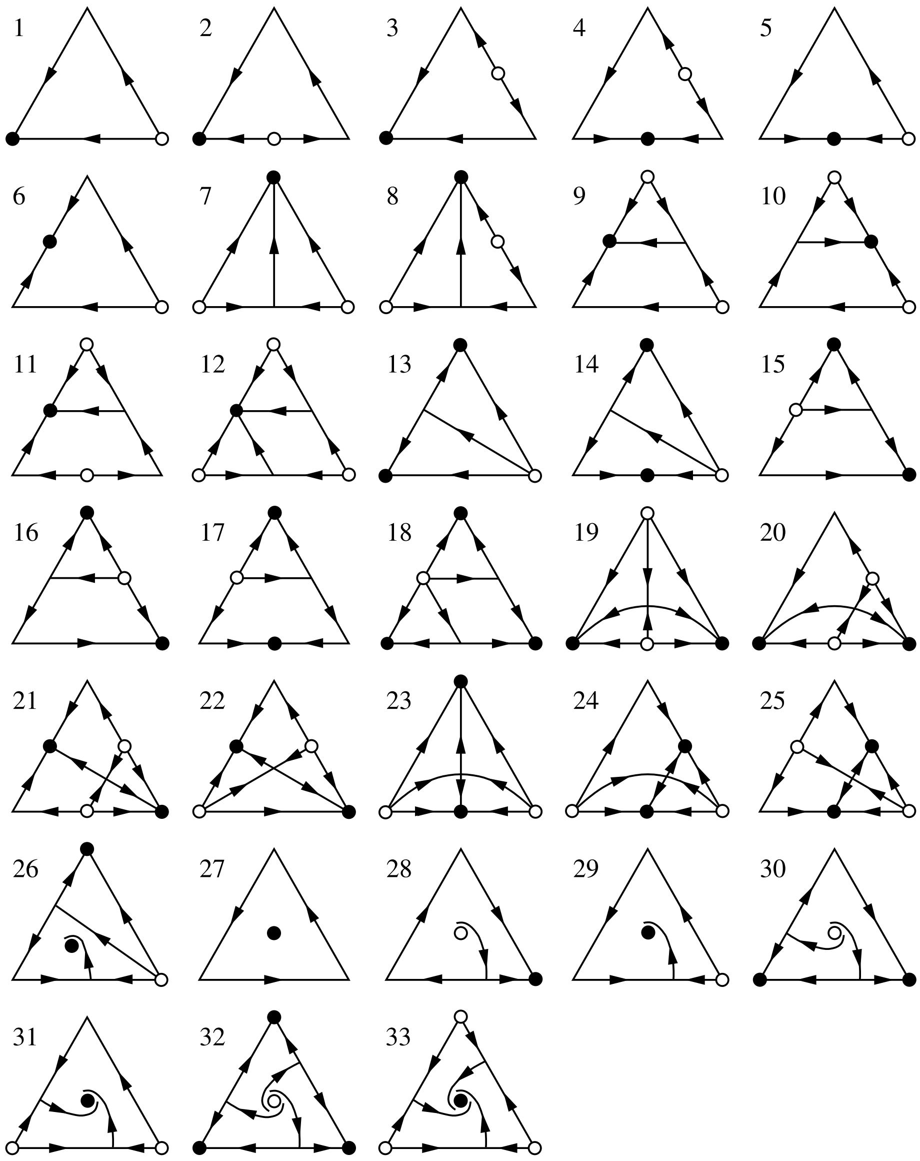


Figure 3. *Carrying simplex dynamics (without periodic orbits) in the nullcline equivalence classes of three-dimensional competitive Lotka–Volterra systems. Redrawn from* [24]*. An attracting equilibrium is denoted by a closed bullet, a repelling equilibrium by an open bullet, and a saddle equilibrium by the intersection of its stable and unstable manifolds.*

the S, I plane are governed by the basic reproduction number $\mathcal{R}_{01}$ ((7) and (8)). In Figure 4, we fix $\tau = 1$ and show numerically how the dynamics on the carrying simplex of system (11) evolve as the disease is strengthened by increasing $\mathcal{R}_{01}$. Specifically, λ is held fixed at $\lambda = 5$, and $r_I = 12, 13, 14, 15$ in the four systems illustrated.

When $r_I = 12$ (Figure 4, top left), then $\mathcal{R}_{01} < 1$, and the DFE at V_S is globally asymptotically stable in $\text{int}\mathbf{R}^3_+$. So with $\lambda = 5$ and $r_I = 12$, the disease is not strong enough to

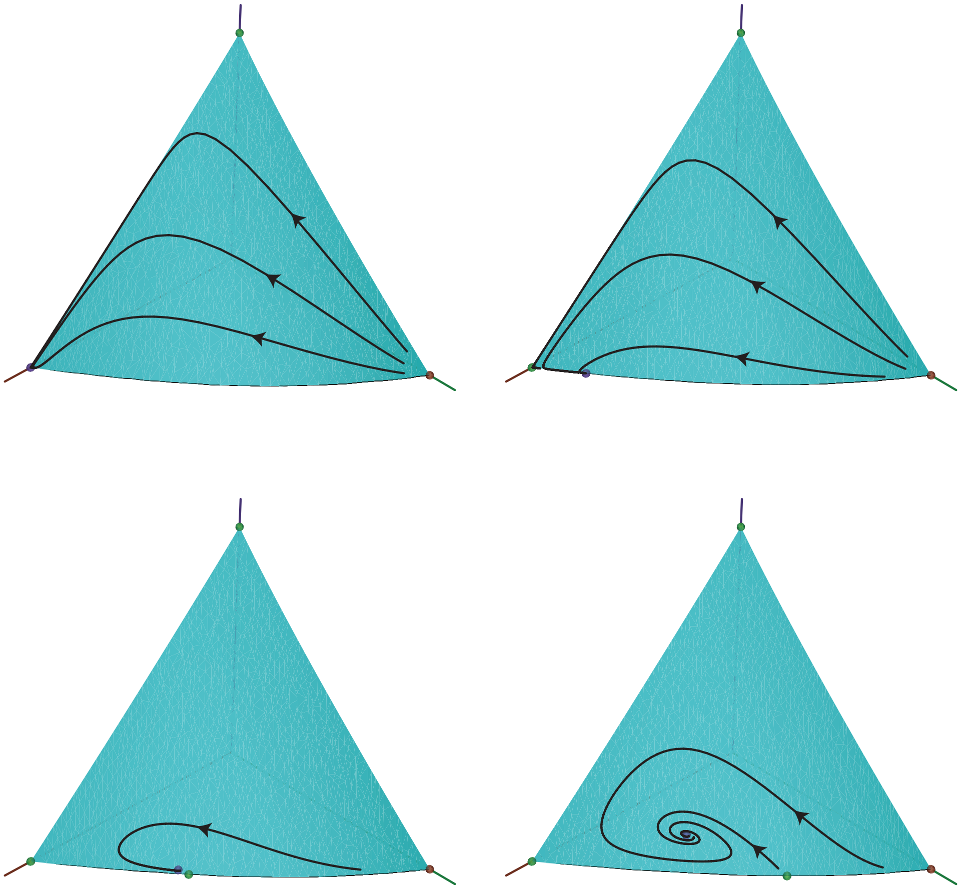


Figure 4. *Dynamics on the carrying simplex of system* (11)*, viewed from above, with* $\tau = 1$*,* $\lambda = 5$*. Top left:* $r_I = 12$*. Top right:* $r_I = 13$*. Bottom left:* $r_I = 14$*. Bottom right:* $r_I = 15$*. The color coding of the axes is* S*: red (left),* I*: green (right), and* N_2*: blue (vertical). Equilibria are indicated by solid dots and are color coded according to their local dynamics on the carrying simplex by repelling: red, saddle: green, attracting: blue.*

affect the long-term demographics. This system lies in class 1 of Figure 3. When $r_I = 12.5$, $\mathcal{R}_{01} = 1$, and there is a bifurcation in which V_S loses stability to the endemic equilibrium, Q_{SI}, as it becomes biologically feasible. When r_I is increased to 13 (Figure 4, top right), $\mathcal{R}_{01} > 1$, and Q_{SI} is globally asymptotically stable in $\text{int}\mathbf{R}^3_+$. This system lies in class 5 of Figure 3. Thus with $r_I = 13$, the disease is strong enough to be endemic in species 1, but species 1 remains strong enough to drive species 2 to extinction. As r_I is increased to 14, there is another bifurcation, in which Q_{SI} loses stability to P as P becomes biologically feasible (Figure 4, bottom left). Thus with $r_I = 14$, the disease has weakened species 1 sufficiently for species 2 to barely survive. As r_I continues to increase to 15 (Figure 4, bottom left), species 1 continues to weaken, and the surviving population of species 2 grows.

Figure 5 shows the dynamics numerically when r_I is increased to 16, and the time scale

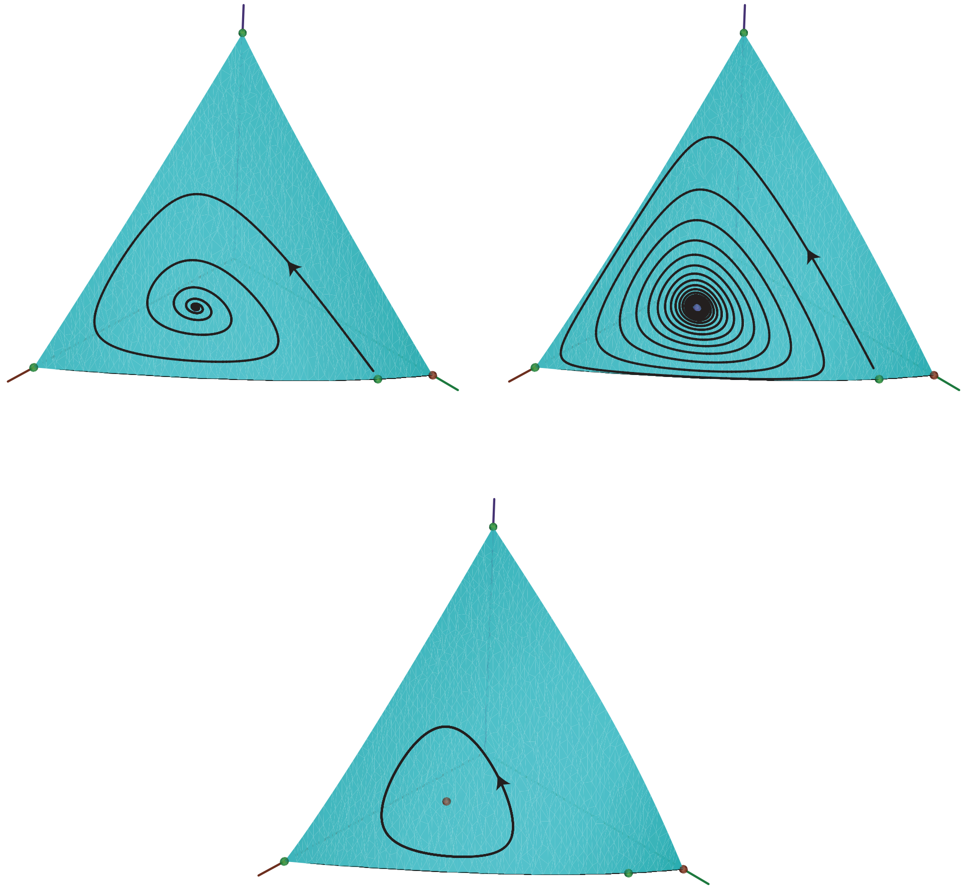


Figure 5. *Dynamics on the carrying simplex of system* (11)*, viewed from above, with* $\lambda = 5$*,* $r_I = 16$*. Top left:* $\tau = 1$*. Top right:* $\tau = 2$*. Bottom:* $\tau = 3$*. The color coding of the axes is* S*: red (left),* I*: green (right), and* N_2*: blue (vertical). Equilibria are indicated by solid dots and are color coded according to their local dynamics on the carrying simplex by repelling: red, saddle: green, attracting: blue.*

of the N_2 equation is changed by increasing τ, representing a faster response by N_2 to its environment. The systems with $r_I = 14, 15, 16$ all lie in class 29 of Figure 3. When $\tau = 1$ (Figure 5, top left), trajectories spiral into P. When τ is increased to 2 (Figure 5, top right), the trajectories spiral more slowly into P, as a pair of eigenvalues of the linearization at P approach the imaginary axis from the left half plane. When τ is further increased to 3 (Figure 5, bottom), this pair of eigenvalues has crossed the imaginary axis, and a Hopf bifurcation has occurred, giving rise to oscillatory endemic coexistence of the species.

Intuitively, if initially the population consists mainly of susceptible individuals of species 1, then the trajectory is drawn toward the endemic equilibrium Q_{SI}, and the number of infectives increases. But species 2 can successfully invade infectives, so species 2 increases and the number of infectives decreases. This leads to a mixed population of mainly species 2 and susceptibles, in which the susceptibles dominate the competition, completing the cycle. Note that by [25, Theorem 4.1], periodic orbits in system (3) with inequalities (4) always have the orientation described here, i.e., counterclockwise as viewed in Figure 5.

In the rest of this section, we prove that much of the behavior illustrated by Example A is true for a general system (3) in Case A. The following theorem shows that if $\mathcal{R}_{01} < 1$, then the disease has no impact on the long-term demographics: species 1 survives without disease and drives species 2 to extinction (as in Figure 4, top left).

Theorem 6.1. *For system* (3) *with inequalities* (4) *and* (10)*, if* $\mathcal{R}_{01} < 1$*, then the DFE,* V_S*, is globally asymptotically stable in* $\mathrm{int}\mathbf{R}^3_+$*.*

Proof. The I nullcline intersects the S and I axes below the S nullcline, as in the proof of Lemma 4.1. The I and S nullclines intersect the N_2 axis at r_I/a_{12} and r_S/a_{12}, respectively, and $r_I/a_{12} < r_S/a_{12}$ since $r_I < r_S$. Thus the I nullcline lies entirely below the S nullcline. By Lemma 5.1, the susceptible population drives the infective population to extinction, and all trajectories in $\mathrm{int}\mathbf{R}^3_+$ are asymptotic to those in the S, N_2 plane. The dynamics in the S, N_2 plane are simply those of Case A, by inequalities (10). Hence species 2 is driven to extinction, and V_S is globally asymptotically stable. ■

Note that when $\mathcal{R}_{01} > 1$, system (3) with inequalities (4) and (10) has no local attractors in the plane $I = 0$, and so the disease persists. Theorem 6.2 shows that as $\mathcal{R}_{01}$ passes through 1, there is a bifurcation at V_S, giving rise to the globally attracting equilibrium Q_{SI}, as in Figure 4, top right.

Theorem 6.2. *Given system* (3) *with inequalities* (4) *and* (10)*,* $\exists\ \epsilon > 0$ *such that if* $\mathcal{R}_{01} \in (1, 1+\epsilon)$*, then* Q_{SI} *is globally asymptotically stable in* $\mathrm{int}\mathbf{R}^3_+$*.*

Proof. Recall from section 5 that system (3) has an equilibrium at $P = (S^*, I^*, N_2^*)$. If $\mathcal{R}_{01} = 1$, then since $r_I = b_I - d_I$, it follows from (7) and (10) that

$$\lambda r_S = a_{11}(r_S - r_I) > \lambda r_2 \frac{a_{11}}{a_{21}}. \tag{12}$$

From (9) this implies that $N_2^* < 0$. Thus, by continuity of $\mathcal{R}_{01}$ in each parameter, $\exists\ \epsilon > 0$ such that for $\mathcal{R}_{01} \in [1, 1+\epsilon)$, $N_2^* < 0$, and so $P \notin \mathrm{int}\mathbf{R}^3_+$.

The linear stability of system (3) about Q_{SI} is determined by the eigenvalues of the Jacobian matrix

$$J = \begin{pmatrix} -a_{11}S & -(a_{11}+\lambda)S & -a_{12}S \\ -(a_{11}-\lambda)I & -a_{11}I & -a_{12}I \\ 0 & 0 & r_2 - a_{21}S - a_{21}I \end{pmatrix}.$$

Consider the upper left 2×2 submatrix $\hat{J}$ of J. Since $a_{11} > \lambda$, the entries of $\hat{J}$ are all negative. Thus, by the Perron–Frobenius theorem [12, p. 182], $\hat{J}$ has a negative eigenvalue, and hence both the eigenvalues of $\hat{J}$ are negative, since the determinant $(\lambda^2 SI)$ is positive. The third eigenvalue of J is $\mu = r_2 - a_{21}S - a_{21}I$. Substituting the values of S and I at Q_{SI} gives $\mu = a_{22}N_2^* < 0$. Thus Q_{SI} is a local attractor.

We now verify that Q_{SI} is the only local attractor, so this system lies in class 5 or 9 of Figure 3, and hence Q_{SI} is globally attracting. By inequalities (10), there is no equilibrium Q_{SN_2}, and V_S is globally attracting in the S, N_2 plane. Thus V_{N_2} is not a local attractor. Since $\mathcal{R}_{01} > 1$, Q_{SI} is globally attracting in the S, I plane (by Lemma 4.2), and so V_S and V_I

are not local attractors. Finally, if Q_{IN_2} is biologically feasible, then it lies on the I nullcline, below the S nullcline (as in the proof of Theorem 6.1), and hence is not locally attracting (see [24, p. 199]). ■

The following theorem shows that as $\mathcal{R}_{01}$ continues to increase and N_2^* passes from negative to positive, P passes through Q_{SI}, becoming biologically feasible, and Q_{SI} loses stability; see Figure 4, bottom and Figure 5. Theorems 6.4 and 6.5 show how the local behavior at P depends on the sign of Δ (see (2)). When $\Delta < 0$, Hopf bifurcations can occur (giving rise to isolated periodic orbits since $r_I \neq r_S$ [6]), whereas when $\Delta > 0$, Hopf bifurcations do not occur. Note that $\Delta < 0$ in Example A.

Theorem 6.3. *Given system* (3) *with inequalities* (4) *and* (10), *Q_{SI} is unstable and P is biologically feasible if and only if $N_2^* > 0$.*

Proof. By the proof of Theorem 6.2, the Jacobian matrix J at Q_{SI} has a positive eigenvalue (so that Q_{SI} is unstable) if and only if $N_2^* > 0$. As remarked in section 5, if $N_2^* > 0$, then $S^* > 0$. If $\Delta \leq 0$, then $I^* > 0$ by using the second inequality of (10). If $\Delta > 0$, then the first inequality of (10) and $N_2^* > 0$ imply that $I^* > 0$. Thus P is biologically feasible. ■

Theorem 6.4. *Given system* (3) *with inequalities* (4) *and* (10), *if P is biologically feasible and $\Delta > 0$, then P is locally asymptotically stable.*

Proof. Linearizing system (3) about P gives a Jacobian matrix with characteristic equation

$$z^3 + z^2(a_{11}(S^* + I^*) + a_{22}N_2^*) + z(\Delta(S^* + I^*)N_2^* + \lambda^2 S^* I^*) + \lambda^2 a_{22} S^* I^* N_2^* = 0.$$

By the Routh–Hurwitz conditions [19, Appendix 2], all roots have negative real parts, and thus P is linearly stable. ■

For a given system (3) satisfying inequalities (4) and (10), each system in the one-parameter family defined below in (13) also satisfies (3), (4), and (10). Thus the Hopf bifurcation described in Theorem 6.5 occurs within Case A of system (3).

Theorem 6.5. *Given system* (3) *with inequalities* (4) *and* (10), *if P is biologically feasible and $\Delta < 0$, then the family of systems*

$$\begin{aligned} S' &= S(r_S - a_{11}S - (a_{11} + \lambda)I - a_{12}N_2), \\ I' &= I(r_I - (a_{11} - \lambda)S - a_{11}I - a_{12}N_2), \\ N_2' &= \tau N_2(r_2 - a_{21}S - a_{21}I - a_{22}N_2) \end{aligned} \tag{13}$$

with $\tau \in (0, \infty)$ admits a Hopf bifurcation.

Proof. Since P is biologically feasible, $S^*, I^*, N_2^* > 0$. Consider the diagonal linear change of coordinates $(x_1, x_2, x_3) = (S/S^*, I/I^*, N_2/N_2^*)$. Then, as in [24, Proposition 3.1], for each $\tau > 0$ the system

$$\begin{aligned} x_1' &= x_1(r_S - a_{11}S^*x_1 - (a_{11} + \lambda)I^*x_2 - a_{12}N_2^*x_3), \\ x_2' &= x_2(r_I - (a_{11} - \lambda)S^*x_1 - a_{11}I^*x_2 - a_{12}N_2^*x_3), \\ x_3' &= \tau x_3(r_2 - a_{21}S^*x_1 - a_{21}I^*x_2 - a_{22}N_2^*x_3) \end{aligned} \tag{14}$$

is topologically equivalent to system (13) and has an equilibrium at $(1, 1, 1)$ since $P = (S^*, I^*, N_2^*)$ is an equilibrium of system (3). The principal 2×2 minors M_{jk} of the coefficient matrix of system (14) are given by

$$M_{12} = \lambda^2 S^* I^* > 0, \qquad M_{23} = \Delta\tau I^* N_2^* < 0, \qquad M_{13} = \Delta\tau S^* N_2^* < 0.$$

Since these minors are not all of the same sign, the result follows from [24, Theorem 3.14 and proof]. ■

7. Case B. For this section, assume that, in the absence of disease, there is initial-condition–dependent competitive exclusion, as in Case B of Figure 1 and Table 1. Introduction of disease is then modeled by system (3) with inequalities (4) and

$$r_2 a_{11} < r_S a_{21} \quad \text{and} \quad r_S a_{22} < r_2 a_{12}. \tag{15}$$

We begin with a numerical example.

Example B. Consider the system

$$\begin{aligned} S' &= S(25 - 10S - 15I - 17N_2),\\ I' &= I(r_I - 5S - 10I - 17N_2),\\ N_2' &= \tau N_2(20 - 8.5S - 8.5I - 12N_2). \end{aligned} \tag{16}$$

In Figure 6 we fix $\lambda = 5$, $\tau = 1$ and strengthen the disease by increasing r_I from 12 to 14. For $r_I = 14$, we then change the time scale τ of species 2 to 1.8. When $r_I = 12$ (Figure 6, top left), $\mathcal{R}_{01} < 1$ and there is initial-condition–dependent competitive exclusion between V_S and V_{N_2} (class 13 in Figure 3). At $r_I = 13$ (Figure 6, top right), $\mathcal{R}_{01} > 1$ and V_S has lost its stability to Q_{SI}. There is now initial-condition–dependent competitive exclusion between the fixed points Q_{SI} and V_{N_2} (class 14 in Figure 3). Increasing r_I to 14 (Figure 6, bottom left), P passes through Q_{SI}, becoming biologically feasible, and the disease weakens species 1 sufficiently for species 2 to survive. Endemic coexistence is now possible for some (but not all) initial conditions (class 26 in Figure 3). Changing the time scale τ (Figure 6, bottom right), P undergoes a Hopf bifurcation, and there is now initial-condition–dependent extinction of species 1 or oscillatory endemic coexistence. Roughly speaking, in the region of the carrying simplex in which N_1 dominates when $\mathcal{R}_{01} < 1$, the sequence of qualitative behavior as $\mathcal{R}_{01}$ increases is the same as that in Case A (section 6), while in the region in which N_2 dominates when $\mathcal{R}_{01} < 1, N_2$ dominates for all $\mathcal{R}_{01}$.

Now consider the general system (3) in Case B. The following theorem shows that, as in Case A, if $\mathcal{R}_{01} < 1$, then the disease has no impact on the long-term demographics.

Theorem 7.1. *For system* (3) *with inequalities* (4) *and* (15)*, if* $\mathcal{R}_{01} < 1$*, then almost every trajectory in* $\mathrm{int}\mathbf{R}^3_+$ *is attracted to* V_S *or* V_{N_2}.

Proof. First note that if $\mathcal{R}_{01} < 1$, then for system (3) with inequalities (4), the I nullcline lies entirely below the S nullcline, as in the proof of Theorem 6.1. Therefore, by Lemma 5.1, all trajectories in $\mathrm{int}\mathbf{R}^3_+$ are asymptotic to those in the S, N_2 plane, where there is initial-condition–dependent competitive exclusion by inequalities (15). Thus, every trajectory, except on the one- or two-dimensional stable manifolds of the other fixed points, is attracted to V_S or V_{N_2}. ■

Theorem 7.2 shows that as $\mathcal{R}_{01}$ passes through 1 and Q_{SI} passes through V_S, becoming biologically feasible, V_S transfers its local stability to Q_{SI} (as in Case A). Thus, for some initial conditions, there is endemic disease in species 1.

Theorem 7.2. *For system* (3) *with inequalities* (4) *and* (15)*, there exists* $\epsilon > 0$ *such that if* $\mathcal{R}_{01} \in (1, 1+\epsilon)$*, then almost every trajectory in* $\mathrm{int}\mathbf{R}^3_+$ *is attracted to* Q_{SI} *or* V_{N_2}.

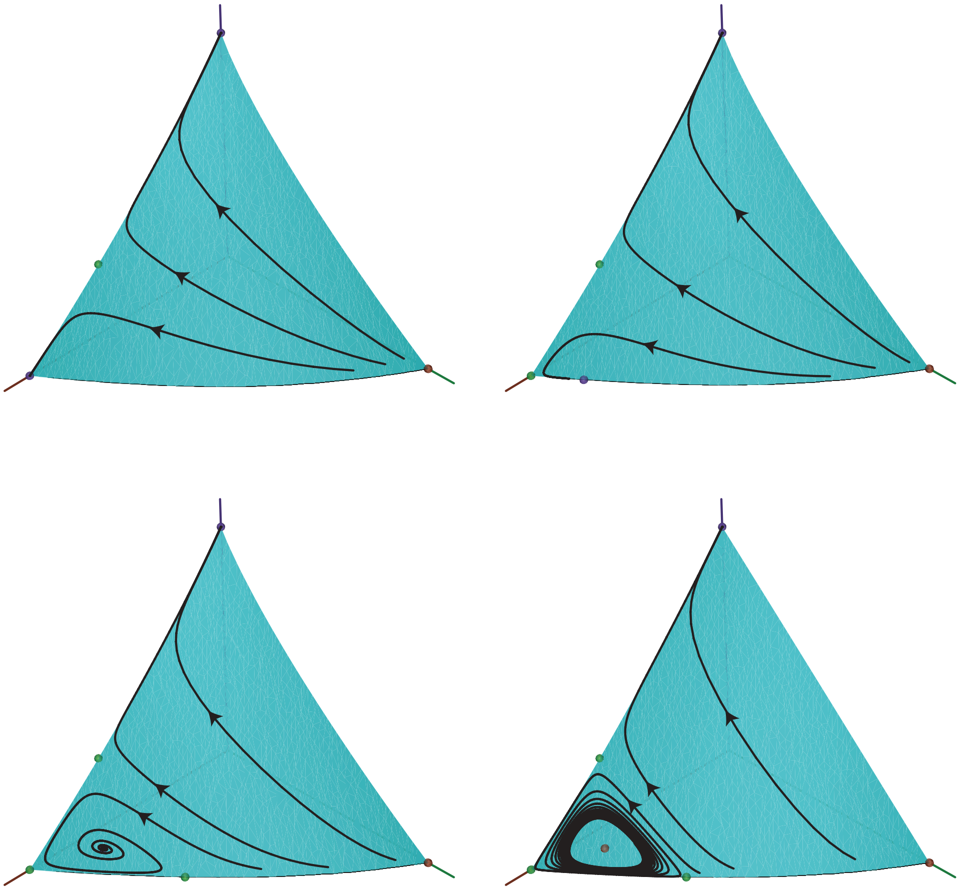


Figure 6. *Dynamics on the carrying simplex of system* (16)*, viewed from above, with $\lambda = 5$. Top left: $r_I = 12$, $\tau = 1$. Top right: $r_I = 13$, $\tau = 1$. Bottom left: $r_I = 14$, $\tau = 1$. Bottom right: $r_I = 14$, $\tau = 1.8$. The color coding of the axes is S: red (left), I: green (right), and N_2: blue (vertical). Equilibria are indicated by solid dots and are color coded according to their local dynamics on the carrying simplex by repelling: red, saddle: green, attracting: blue.*

Proof. As in the proof of Theorem 6.2, $N_2^* < 0$; thus $P \notin \text{int}\mathbf{R}^{\mathbf{3}}_{+}$. It is then easy to verify that the only local attractors are Q_{SI} and V_{N_2}, and this system lies in class 14 of Figure 3. ■

Theorem 7.3 shows that as $\mathcal{R}_{01}$ increases further, the qualitative behavior differs from that in Case A, in the sense that every family of systems (13) in Case B that admits stable endemic coexistence also admits oscillatory endemic coexistence. See Figure 6, bottom right.

Theorem 7.3. *Consider system* (3) *with inequalities* (4) *and* (15)*.*

(i) *Q_{SI} is unstable if and only if $N_2^* > 0$.*

(ii) *There exists $\epsilon > 0$ such that if $N_2^* \in (0, \epsilon)$, then P is biologically feasible.*

(iii) *If P is biologically feasible, then every family of systems* (13) *admits a Hopf bifurcation.*

Proof. Statement (i) is proved as in Theorem 6.3. If $N_2^* > 0$, then $S^* > 0$. If $N_2^* = 0$, then $P = Q_{SI}$, and so by continuity, there exists $\epsilon > 0$ such that $I^* > 0$ if $N_2^* \in [0, \epsilon)$, thereby proving statement (ii). Statement (iii) is proved as in Theorem 6.5, since $\Delta < 0$ by inequalities (15). ■

8. Case C.

For this section, assume that in the absence of disease, species 1 and 2 coexist, as in Case C of Figure 1 and Table 1. Introduction of disease is then modeled by system (3) with inequalities (4) and

$$r_S a_{21} < r_2 a_{11} \quad \text{and} \quad r_2 a_{12} < r_S a_{22}. \tag{17}$$

We begin with a numerical example.

Example C. Consider the system

$$\begin{aligned} S' &= S(20 - 15S - (15+\lambda)I - 10N_2), \\ I' &= I(15 - (15-\lambda)S - 15I - 10N_2), \\ N_2' &= N_2(15 - 10S - 10I - 20N_2), \end{aligned} \tag{18}$$

in which we fix $r_I = 15$ and strengthen the disease by increasing λ from 3.5 to 4.5.

When $\lambda = 3.5$ (Figure 7, top left), $\mathcal{R}_{01} < 1$ and the DFE at Q_{SN_2} is globally attracting; see class 9 of Figure 3. When $\lambda = 3.5$ (Figure 7, top right), $\mathcal{R}_{01} > 1$, Q_{SI} passes through V_S, becoming biologically feasible, but the DFE at Q_{SN_2} remains globally attracting; see class 12 of Figure 3. Thus, in this case, $\mathcal{R}_{01} = 1$ is not a threshold for endemic disease. When $\lambda = 4$, P passes through Q_{SN_2}, and so at $\lambda = 4.1$ and $\lambda = 4.5$ (Figure 7, bottom), P is biologically feasible and globally attracting; see class 33 of Figure 3, and [20]. Thus species 1 and species 2 continue to stably coexist, but with endemic disease in species 1.

Now consider the general system (3) in Case C, and note that $\Delta > 0$ by inequalities (17). By contrast to Cases A and B, a basic reproduction number $\mathcal{R}_0$ for the full model, rather than $\mathcal{R}_{01}$, acts as a disease threshold. This is because, in the absence of disease, the two species coexist. Hence Q_{SN_2} (rather than V_S) is the DFE for Case C. Using the next generation matrix [21] of system (3) at Q_{SN_2}, the basic reproduction number $\mathcal{R}_0$ is defined by

$$\mathcal{R}_0 = \frac{\lambda S + b_I}{a_{11}S + a_{12}N_2 + d_I} = \frac{(\lambda/\Delta)(r_S a_{22} - r_2 a_{12}) + b_I}{r_S + d_I}. \tag{19}$$

This parameter has the same interpretation as $\mathcal{R}_{01}$ (see section 4), except that species 2 is now present (but is not infected by the disease). Note that in Example C, $\lambda = 4$ gives $\mathcal{R}_0 = 1$.

Theorem 8.1. *For system* (3) *with inequalities* (4) *and* (17)*, if* $\mathcal{R}_0 < 1$*, then the DFE,* Q_{SN_2}*, is globally asymptotically stable in* $\text{int}\mathbf{R}^3_+$.

Proof. Recall that $r_I = b_I - d_I$. Thus, by (9), if $\mathcal{R}_0 < 1$, then $I^* < 0$, and so $P \notin \text{int}\mathbf{R}^3_+$. Also, since $\mathcal{R}_0 < 1$, the Jacobian matrix at Q_{SN_2} of system (3) with inequalities (17) has all negative eigenvalues. Thus Q_{SN_2} is locally attracting. By inequalities (17), V_S and V_{N_2} are not locally attracting in the S, N_2 plane and hence are not locally attracting in $\mathbf{R}^3_+$. The I nullcline lies below the S nullcline on the I, N_2 plane, by inequalities (4), and hence V_I and Q_{IN_2} (if biologically feasible) are not locally attracting [24, p. 199]. Finally, if Q_{SI} is biologically feasible, then by inequalities (17) it lies on the S nullcline, below the N_2 nullcline,

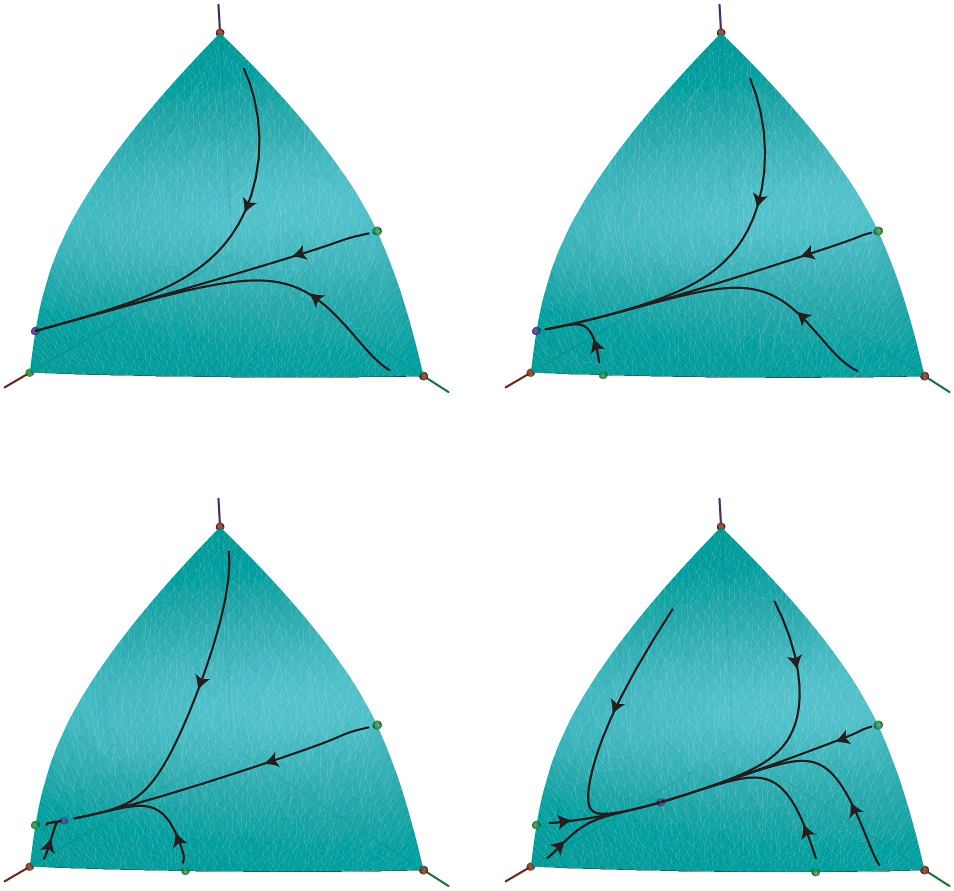


Figure 7. *Dynamics on the carrying simplex of system* (18)*, viewed from above, with* $r_I = 15$*. Top left:* $\lambda = 3.5$*. Top right:* $\lambda = 3.9$*. Bottom left:* $\lambda = 4.1$*. Bottom right:* $\lambda = 4.5$*. The color coding of the axes is* S*: red (left),* I*: green (right), and* N_2*: blue (vertical). Equilibria are indicated by solid dots and are color coded according to their local dynamics on the carrying simplex by repelling: red, saddle: green, attracting: blue.*

and hence is not locally attracting [24, p. 199]. The system is therefore in class 6, 9, or 12 of Figure 3, and hence Q_{SN_2} is globally attracting in int$\mathbf{R}^3_+$. ■

By inequalities (17), $\mathcal{R}_0 < \mathcal{R}_{01}$. Thus Theorem 8.1 also shows that if $\mathcal{R}_{01} < 1$, then Q_{SN_2} is globally asymptotically stable in int$\mathbf{R}^3_+$. If $\mathcal{R}_0 > 1$, then $I^* > 0$, and by inequalities (17), $N_2^* > 0$ and $S^* > 0$, since $\mathcal{R}_{01} > 1$. In other words, as $\mathcal{R}_0$ increases through 1, P passes through Q_{SN_2}, becoming biologically feasible. By the Routh–Hurwitz conditions, P is then locally asymptotically stable. In [20] periodic orbits are ruled out using a generalized Bendixson–Dulac criterion [5] on the carrying simplex. This result is now applied to find conditions under which P is globally asymptotically stable. We conjecture that the hypotheses of Theorem 8.2 can be weakened by dropping inequality (20).

Theorem 8.2. *For system* (3) *with inequalities* (4) *and* (17), *if* $\mathcal{R}_0 > 1$ *and*

$$r_2 a_{12} < r_I a_{22}, \tag{20}$$

then the coexisting endemic equilibrium P *is globally asymptotically stable in* $\operatorname{int}\mathbf{R}^3_+$.

Proof. Since $\mathcal{R}_0 > 1$, $P \in \operatorname{int}\mathbf{R}^3_+$. By [20, Corollary 1.4], it is sufficient to prove that V_S, V_I, V_{N_2} are all local repellors of the flow restricted to the carrying simplex, Σ. By inequality $r_S a_{21} < r_2 a_{11}$ of (17) and $\mathcal{R}_{01} > 1$, V_S lies below the I and N_2 nullclines on the S axis and hence is locally repelling on Σ [24, p. 199]. Similarly, by $(a_{11} + \lambda)r_I < a_{11} r_S$ of (4) and $r_S a_{21} < r_2 a_{11}$ of (17), V_I is locally repelling on Σ. Finally, by $r_2 a_{12} < r_S a_{22}$ of (17) and (20), V_{N_2} is locally repelling on Σ. ■

9. Case D.

For this section, assume that, in the absence of disease, species 2 drives species 1 to extinction, as in Case D of Figure 1 and Table 1. Introduction of disease is then modeled by system (3) with inequalities (4) and

$$r_S a_{21} < r_2 a_{11} \quad \text{and} \quad r_S a_{22} < r_2 a_{12} \tag{21}$$

from Table 1, Case D. Theorem 9.1 shows, not surprisingly, that weakening species 1 by the introduction of disease does not change the long-term demographics in this case: species 1 is always driven to extinction.

Theorem 9.1. *For system* (3) *with inequalities* (4) *and* (21), V_{N_2} *is globally asymptotically stable.*

Proof. By inequalities (21), the S nullcline lies entirely below the N_2 nullcline in $\operatorname{int}\mathbf{R}^3_+$. So, by Lemma 5.1, species N_2 drives the susceptible population to extinction, and all trajectories in $\operatorname{int}\mathbf{R}^3_+$ are asymptotic to those in the I, N_2 plane. Now in the I, N_2 plane, the I nullcline lies entirely below the N_2 nullcline (using inequalities (21) and $r_I < r_S$), and so by Lemma 5.1 again, species 2 drives the infective population to extinction. Thus V_{N_2} is globally asymptotically stable. ■

Acknowledgments. We are grateful to Dmitry Gokhman, M. Drew LaMar, and David McCobb for helpful conversations and technical assistance.

REFERENCES

[1] R. M. Anderson and R. M. May, *The invasion, persistence and spread of infectious diseases within animal and plant communities*, Phil. Trans. Roy. Soc. London B, 314 (1986), pp. 553–570.

[2] M. Begon and R. G. Bowers, *Beyond host pathogen dynamics*, in Ecology of Infectious Diseases in Natural Populations, B. T. Grenfell and A. P. Dobson, eds., Cambridge University Press, Cambridge, UK, 1995, pp. 478–509.

[3] R. G. Bowers and J. Turner, *Community structure and interplay between interspecific infection and competition*, J. Theor. Biol., 187 (1997), pp. 95–109.

[4] S. Busenberg and K. L. Cooke, *Vertically Transmitted Diseases*, Biomath. 23, Springer-Verlag, New York, 1993.

[5] S. Busenberg and P. van den Driessche, *A method for proving the non-existence of limit cycles*, J. Math. Anal. Appl., 172 (1993), pp. 463–479.

[6] J. Coste, J. Peyraud, and P. Coullet, *Asymptotic behaviors in the dynamics of competing species*, SIAM J. Appl. Math., 36 (1979), pp. 516–543.

[7] L. Q. Gao and H. W. Hethcote, *Disease transmission models with density-dependent demographics*, J. Math. Biol., 30 (1992), pp. 717–731.

[8] J. V. Greenman and P. J. Hudson, *Infected coexistence instability with and without density-dependent regulation*, J. Theor. Biol., 185 (1997), pp. 345–356.

[9] J. V. Greenman and P. J. Hudson, *Host exclusion and coexistence in apparent and direct competition: An application of bifurcation theory*, Theor. Pop. Biol., 56 (1999), pp. 48–64.

[10] H. W. Hethcote and S. A. Levin, *Periodicity in epidemiological models*, in Applied Mathematical Ecology, S. A. Levin, T. G. Hallam, and L. J. Gross, eds., Springer-Verlag, New York, 1999, pp. 193–211.

[11] M. W. Hirsch, *Systems of differential equations that are competitive or cooperative.* III: *Competing species*, Nonlinearity, 1 (1988), pp. 51–71.

[12] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.

[13] J. Hofbauer and J. W.-H. So, *Multiple limit cycles for three dimensional competitive Lotka–Volterra equations*, Appl. Math. Lett., 7 (1994), pp. 65–70.

[14] R. D. Holt and J. Pickering, *Infectious diseases and species coexistence: A model of Lotka-Volterra form*, Am. Nat., 126 (1985), pp. 196–211.

[15] M. D. Lamar and M. L. Zeeman, *CSimplex, a Geomview Module for Visualizing the Carrying Simplex of a Competitive Lotka-Volterra System*, software to be available.

[16] S. Levy, T. Munzner, M. Phillips, et al., *Geomview*, The Geometry Center, Minneapolis, MN, 1996, http://www.geom.umn.edu.

[17] Z. Lu and Y. Luo, *Two limit cycles in three-dimensional Lotka-Volterra systems*, Comput. Math. Appl., 44 (2002), pp. 51–66.

[18] F. Montes de Oca and M. L. Zeeman, *Extinction in nonautonomous competitive Lotka-Volterra systems*, Proc. Amer. Math. Soc., 124 (1996), pp. 3677–3687.

[19] J. D. Murray, *Mathematical Biology*, Biomath. 19, Springer-Verlag, New York, 1993.

[20] P. van den Driessche and M. L. Zeeman, *Three-dimensional competitive Lotka–Volterra systems with no periodic orbits*, SIAM J. Appl. Math., 58 (1998), pp. 227–234.

[21] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), pp. 29–48.

[22] E. Venturino, *The effects of diseases on competing species*, Math. Biosci., 174 (2001), pp. 111–131.

[23] D. Xiao and W. Li, *Limit cycles for the competitive three dimensional Lotka-Volterra systems*, J. Differential Equations, 164 (2000), pp. 1–15.

[24] M. L. Zeeman, *Hopf bifurcations in competitive three dimensional Lotka–Volterra systems*, Dyn. Stab. Syst., 8 (1993), pp. 189–217.

[25] M. L. Zeeman, *On directed periodic orbits in three-dimensional competitive Lotka–Volterra systems*, in Differential Equations and Applications to Biology and to Industry, M. Martelli et al., eds., World Scientific, River Edge, NJ, 1996, pp. 563–572.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 4, pp. 620–635

Dynamic Bifurcation of the Ginzburg–Landau Equation[*]

Tian Ma[†], Jungho Park[‡], and Shouhong Wang[‡]

Abstract. We study in this article the bifurcation and stability of the solutions of the Ginzburg–Landau equation, using a notion of bifurcation called attractor bifurcation. We obtain in particular a full classification of the bifurcated attractor and the global attractor as λ crosses the first critical value of the linear problem. Bifurcations from the rest of the eigenvalues of the linear problem are obtained as well.

Key words. Ginzburg–Landau equation, bifurcation, stability

AMS subject classifications. 35, 37

DOI. 10.1137/040603747

1. Introduction.

In this article, we consider the bifurcation of attractors and invariant sets of the complex Ginzburg–Landau (GL) equation, which reads

$$\frac{\partial u}{\partial t} - (\alpha + i\beta)\triangle u + (\sigma + i\rho)|u|^2 u - \lambda u = 0, \tag{1.1}$$

where the unknown function $u : \Omega \times [0, \infty) \to \mathbb{C}$ is a complex-valued function and $\Omega \subset \mathbb{R}^n$ is an open, bounded, and smooth domain in $\mathbb{R}^n$ $(1 \leq n \leq 3)$. The parameters α, β, σ, ρ, and λ are real numbers and

$$\alpha > 0, \quad \sigma > 0. \tag{1.2}$$

The initial condition for (1.1) is given by

$$u(x, 0) = \phi + i\psi. \tag{1.3}$$

Also, (1.1) is supplemented with either the Dirichlet boundary condition,

$$u|_{\partial\Omega} = 0, \tag{1.4}$$

or the periodic boundary condition,

$$\Omega = (0, 2\pi)^n \text{ and } u \text{ is } \Omega\text{-periodic}. \tag{1.5}$$

[*] Received by the editors February 2, 2004; accepted for publication (in revised form) by D. Holm July 1, 2004; published electronically December 27, 2004. The work was supported in part by the Office of Naval Research, by the National Science Foundation, and by the National Science Foundation of China.
http://www.siam.org/journals/siads/3-4/60374.html

[†] Department of Mathematics, Sichuan University, Chengdu, People's Republic of China and Department of Mathematics, Indiana University, Bloomington, IN 47405 (tima@indiana.edu).

[‡] Department of Mathematics, Indiana University, Bloomington, IN 47405 (junjupar@indiana.edu, showang@indiana.edu).

The GL equation is an important equation in a number of scientific fields. It is directly related to the GL theory of superconductivity. In this context, the unknown function is the order parameter, the constants β and ρ are usually zero, and the bifurcation parameter λ is the GL parameter; see [11] and the references therein.

In fluid dynamics the GL equation is found, for example, in the study of Poiseuille flow, the nonlinear growth of convection rolls in the Rayleigh–Beńard problem, and Taylor–Couette flow. In this case, the bifurcation parameter λ plays the role of a Reynolds number. The equation also arises in the study of chemical systems governed by reaction-diffusion equations.

There are extensive studies from the mathematical point of view for the GL equation, and we refer in particular to [2, 3, 1, 5, 6, 7, 11] and the references therein for studies related to the global attractors, inertial manifolds, and soft and hard turbulences described by the GL equations.

We study in this article the bifurcation and stability of the solutions of the complex GL equation. A nonlinear theory for this problem is established in this article using a notion of bifurcation called attractor bifurcation and its corresponding theorem developed recently by the authors in [9, 8]; see [10], a new book by two of the authors. The main objectives of this theory include

(1) existence of bifurcation when the system parameter crosses some critical numbers,
(2) dynamic stability of bifurcated solutions, and
(3) the structure/patterns and their stability and transitions in the physical space.

More precisely, the main theorem associated with the attractor bifurcation states that as the control parameter crosses a certain critical value when there are $m+1$ ($m \geq 0$) eigenvalues across the imaginary axis, the system bifurcates from a trivial steady state solution to an attractor with dimension between m and $m+1$, provided the critical state is asymptotically stable. There are a few important features of the attractor bifurcation. First, the bifurcation attractor does not include the trivial steady state and is stable; hence it is physically important. Second, the attractor contains a collection of solutions of the evolution equation, including possibly steady states and periodic orbits as well as homoclinic and heteroclinic orbits. Third, it provides a unified point of view on dynamic bifurcation and can be applied to many problems in physics and mechanics. Fourth, from the application point of view, the Krasnoselskii–Rabinowitz theorem requires the number of eigenvalues $m+1$ crossing the imaginary axis being an odd integer, and the Hopf bifurcation is for the case where $m+1=2$. However, the new attractor bifurcation theorem obtained can be applied to cases for all $m \geq 0$. In addition, the bifurcated attractor, as mentioned earlier, is stable, which is another subtle issue for other known bifurcation theorems.

For the GL equation, bifurcation is obtained with respect to the parameter λ, and the main results obtained can be summarized as follows.

First, for the GL equation with the Dirichlet boundary condition, let λ_1 be the first eigenvalue of the elliptic operator $-\triangle$. Our main results in this case include the following.

1. If $\lambda \leq \alpha\lambda_1$, the trivial solution $u=0$ is globally asymptotically stable. The global attractor of the GL equation consists exactly of the trivial steady state solution $u=0$.

2. As λ crosses $\alpha\lambda_1$, i.e., there exists an $\epsilon > 0$ such that for any $\alpha\lambda_1 < \lambda < \alpha\lambda_1 + \epsilon$, the GL problem bifurcates from the trivial solution an attractor Σ_λ. The bifurcated attractor Σ_λ

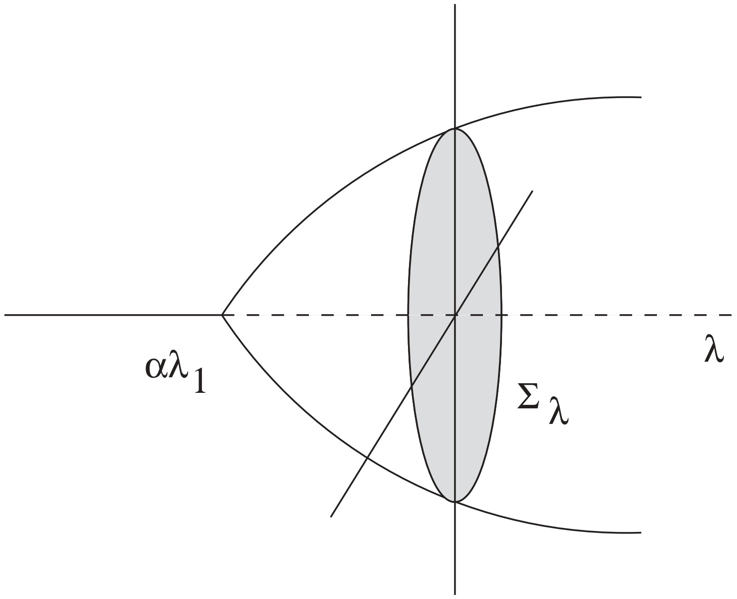


Figure 1. *Bifurcation diagram for the GL equation with the Dirichlet boundary condition: (1) Bifurcation appears at $\lambda = \alpha\lambda_1$; (2) bifurcated attractor $\Sigma_\lambda = S^1$ is the boundary of the shaded region; and (3) the global attractor $\mathcal{A}_\lambda$ is the 2D disk, shown as the shaded region. Here the dotted line stands for the unstable trivial solution $u = 0$.*

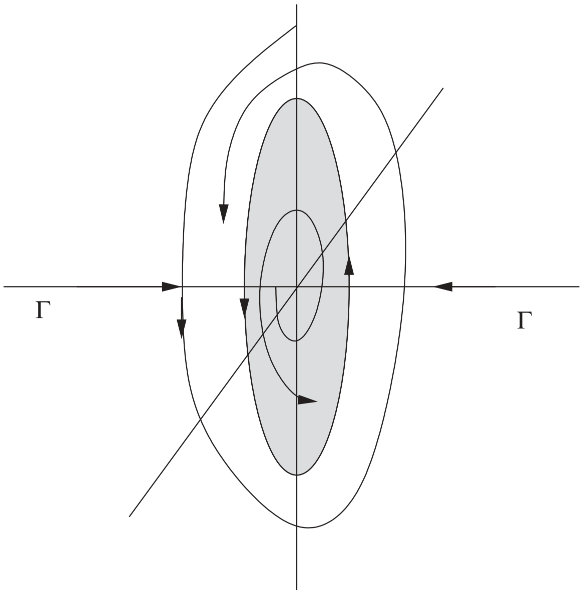


Figure 2. *Phase space structure in $L^2(\Omega, \mathbb{C}) \times \{\lambda\}$ in the case where $|\beta| + |\rho| \neq 0$. Here the bifurcated attractor $\Sigma_\lambda = S^1$ is a stable limiting cycle.*

attracts the open set $L^2(\Omega, \mathbb{C})/\Gamma$, where Γ is the stable manifold of $u = 0$ having codimension two in $L^2(\Omega, \mathbb{C})$.

More detailed structure of this bifurcated attractor can be classified as follows (see Figures 1 and 2).

(a) If $|\beta| + |\rho| \neq 0$, then the bifurcated attractor consists of exactly one stable limiting cycle, i.e., $\Sigma_\lambda = S^1$, which is asymptotically stable. The global attractor $\mathbb{A}_\lambda$ is a two-dimensional (2D) disk consisting of the stable limiting cycle $\Sigma_\lambda = S^1$, the (unstable) trivial steady state solution $u = 0$, and orbits connecting $\Sigma_\lambda = S^1$ and $u = 0$.
In particular, if $\beta \neq 0$, then the bifurcation is a Hopf bifurcation to a stable limiting cycle.

(b) If $\beta = \rho = 0$, then the bifurcated attractor Σ_λ has dimension between 1 and 2 and is a limit of a sequence of 2D annuli M_k with $M_{k+1} \subset M_k$, i.e., $\Sigma_\lambda = \cap_{k=1}^{\infty} M_k$.
Again in this case, the global attractor $\mathbb{A}_\lambda$ is 2D, consisting of Σ_λ, $u = 0$, and the connecting orbits between them.[1]

Second, for the GL equations equipped with the periodic boundary condition, similar results can be obtained as well. In particular, in the case where $|\beta| + |\rho| \neq 0$, we prove that the bifurcated attractor Σ_λ is S^1, containing no steady state solutions, and the global attractor $\mathbb{A}_\lambda$ is a 2D ball consisting of the trivial steady state $u = 0$, Σ_λ, and the orbits connecting them.

Finally, bifurcation from *any* eigenvalue of the Laplacian can also be obtained as for the first eigenvalue. It is worth mentioning that the complete structure of the global attractor for the bifurcations from the first eigenvalue is obtained, while no such information is available for bifurcations from the rest eigenvalues.

Important work on lower and upper bounds of the global attractor of the GL equation, together with their physical mechanisms, was done in the 1980's in [2, 3, 1]. As mentioned earlier, the main objective of this article is to study bifurcation and transitions from the trivial solution. Hence we focus only on the local attractor near the trivial solution, which is part of the global attractor. Of course, near the first eigenvalue, complete information for both the global attractor and the bifurcated attractor is obtained in this article. For λ near other eigenvalues, the results here demonstrate only the transitions of the trivial solution and provide some partial information on the low bounds of the global attractor. As far as the dimension of the global attractor is concerned, our results are consistent with the work in [2, 3, 1].

The paper is organized as follows. In section 2, we recall the attractor bifurcation theory. Sections 3 and 4 study the bifurcation of the GL equations for the Dirichlet boundary condition and for the periodic boundary condition, respectively. Section 5 deals with bifurcation from the rest of the eigenvalues.

2. Abstract bifurcation theory.

2.1. Preliminary. We recall in this section a general theory on attractor bifurcation for nonlinear evolution equations; see [9, 8].

Let H and H_1 be two Hilbert spaces and $H_1 \hookrightarrow H$ be a dense and compact inclusion. We

[1] Using a different method, we can in fact prove that Σ_λ is also homeomorphic to S^1, which shall be reported elsewhere.

consider the nonlinear evolution equations

$$\frac{du}{dt} = L_\lambda u + G(u, \lambda), \tag{2.1}$$

$$u(0) = u_0, \tag{2.2}$$

where $u : [0, \infty) \to H$ is the unknown function, $\lambda \in \mathbb{R}$ is the system parameter, and $L_\lambda : H_1 \to H$ are parameterized linear completely continuous fields depending continuously on $\lambda \in \mathbb{R}^1$, which satisfy

$$\begin{cases} L_\lambda = -A + B_\lambda & \text{is a sectorial operator,} \\ A : H_1 \to H & \text{a linear homeomorphism,} \\ B_\lambda : H_1 \to H & \text{the parameterized linear compact operators.} \end{cases} \tag{2.3}$$

It is easy to see that L_λ generates an analytic semigroup $\{e^{-tL_\lambda}\}_{t\geq 0}$. Then we can define fractional power operators L_λ^α for any $0 \leq \alpha \leq 1$ with domain $H_\alpha = D(L_\lambda^\alpha)$ such that $H_{\alpha_1} \subset H_{\alpha_2}$ if $\alpha_1 > \alpha_2$, and $H_0 = H$.

Furthermore, we assume that the nonlinear terms $G(\cdot, \lambda) : H_\alpha \to H$ for some $1 > \alpha \geq 0$ are a family of parameterized C^r bounded operators ($r \geq 1$) continuously depending on the parameter $\lambda \in \mathbb{R}^1$, such that

$$G(u, \lambda) = o(\|u\|_{H_\alpha}) \quad \forall\ \lambda \in \mathbb{R}^1. \tag{2.4}$$

Actually, in this paper we need only the following conditions for the operator $L_\lambda = -A + B_\lambda$, which ensure that L_λ is a sectorial operator.

Let there be a eigenvalue sequence $\{\rho_k\} \subset \mathbb{C}$ and an eigenvector sequence $\{e_k, h_k\} \subset H_1$ of A:

$$\begin{cases} Az_k = \rho_k z_k,\ z = e_k + ih_k, \\ \operatorname{Re} \rho_k \to +\infty \text{ as } k \to \infty, \\ |\operatorname{Im} \rho_k / (\operatorname{Re} \rho_k + a)| \leq C \text{ for some constants } a, C > 0, \end{cases} \tag{2.5}$$

such that $\{e_k, h_k\}$ is a basis of H.

Condition (2.5) implies that A is a sectorial operator. Hence we can define fractional power operator A^α with domain $H_\alpha = D(A^\alpha)$. Then for the operator $B_\lambda : H_1 \longrightarrow H$, we assume that there is a constant $0 \leq \theta < 1$ such that

$$B_\lambda : H_\theta \longrightarrow H \text{ bounded } \forall\ \lambda \in \mathbb{R}. \tag{2.6}$$

Let $\{S_\lambda(t)\}_{t\geq 0}$ be an operator semigroup generated by (2.1) which enjoys the following properties:

(i) For any $t \geq 0$, $S_\lambda(t) : H \to H$ is a linear continuous operator.

(ii) $S_\lambda(0) = I : H \to H$ is the identity on H.

(iii) For any $t, s \geq 0$, $S_\lambda(t + s) = S_\lambda(t) \cdot S_\lambda(s)$.

Then the solution of (2.1) and (2.2) can be expressed as

$$u(t) = S_\lambda(t)u_0, \qquad t \geq 0.$$

Definition 2.1. *A set $\Sigma \subset H$ is called an invariant set of (2.1) if $S(t)\Sigma = \Sigma$ for any $t \geq 0$. An invariant set $\Sigma \subset H$ of (2.1) is called an attractor if Σ is compact, and there exists a neighborhood $U \subset H$ of Σ such that for any $\varphi \in U$ we have*

$$\lim_{t\to\infty} \operatorname{dist}_H(u(t,\varphi),\Sigma) = 0. \tag{2.7}$$

The largest open set U satisfying (2.7) is called the basin of attraction of Σ.

Definition 2.2.

1. *We say that (2.1) bifurcates from $(u,\lambda) = (0,\lambda_0)$ an invariant set Ω_λ if there exists a sequence of invariant sets $\{\Omega_{\lambda_n}\}$ of (2.1), $0 \notin \Omega_{\lambda_n}$, such that*

$$\lim_{n\to\infty} \lambda_n = \lambda_0,$$
$$\lim_{n\to\infty} \max_{x\in\Omega_{\lambda_n}} |x| = 0.$$

2. *If the invariant sets Ω_λ are attractors of (2.1), then the bifurcation is called an attractor bifurcation.*
3. *If Ω_λ are attractors and are homotopy equivalent to an m-dimensional sphere S^m, then the bifurcation is called an S^m-attractor bifurcation.*

2.2. Center manifold theorems. We assume that the spaces H_1 and H can be decomposed into

$$\begin{cases} H_1 = E_1^\lambda \oplus E_2^\lambda, & \dim E_1^\lambda < \infty, \quad \text{near } \lambda_0 \in \mathbb{R}^1, \\ H = \widetilde{E}_1^\lambda \oplus \widetilde{E}_2^\lambda, & \widetilde{E}_1^\lambda = E_1^\lambda, \quad \widetilde{E}_2^\lambda = \text{closure of } E_2^\lambda \text{ in } H, \end{cases} \tag{2.8}$$

where E_1^λ and E_2^λ are two invariant subspaces of L_λ; i.e., L_λ can be decomposed into $L_\lambda = \mathcal{L}_1^\lambda \oplus \mathcal{L}_2^\lambda$ such that for any λ near λ_0,

$$\begin{cases} \mathcal{L}_1^\lambda = L_\lambda|_{E_1^\lambda} : E_1^\lambda \to \widetilde{E}_1^\lambda, \\ \mathcal{L}_2^\lambda = L_\lambda|_{E_2^\lambda} : E_2^\lambda \to \widetilde{E}_2^\lambda, \end{cases} \tag{2.9}$$

where all eigenvalues of $\mathcal{L}_2^\lambda$ possess negative real parts, and all eigenvalues of $\mathcal{L}_1^\lambda$ possess nonnegative real parts at $\lambda = \lambda_0$.

Thus, for λ near λ_0, (2.1) can be rewritten as

$$\begin{cases} \dfrac{dx}{dt} = \mathcal{L}_1^\lambda x + G_1(x,y,\lambda), \\ \dfrac{dy}{dt} = \mathcal{L}_2^\lambda y + G_2(x,y,\lambda), \end{cases} \tag{2.10}$$

where $u = x+y \in H_1$, $x \in E_1^\lambda$, $y \in E_2^\lambda$, $G_i(x,y,\lambda) = P_i G(u,\lambda)$, and $P_i : H \to \widetilde{E}_i$ are canonical projections. Furthermore, we let

$$E_2^\lambda(\alpha) = E_2^\lambda \cap H_\alpha,$$

with α given by (2.4).

Theorem 2.3 (center manifold theorem, [4]). *Assume (2.4)–(2.6), (2.8), and (2.9). Then there exist a neighborhood of λ_0 given by $|\lambda - \lambda_0| < \delta$ for some $\delta > 0$, a neighborhood $B_\lambda \subset E_1^\lambda$ of $x = 0$, and a C^1 function $h(\cdot,\lambda) : B_\lambda \to E_2^\lambda(\alpha)$, depending continuously on λ, such that*

1. *$h(0,\lambda) = 0$, $D_x h(0,\lambda) = 0$;*
2. *the set*

$$M_\lambda = \left\{ (x,y) \in H_1 \;\middle|\; x \in B_\lambda,\ y = h(x,\lambda) \in E_2^\lambda(\alpha) \right\},$$

called center manifolds, is locally invariant for (2.1); i.e., for any $u_0 \in M_\lambda$,

$$u_\lambda(t,u_0) \in M_\lambda \quad \forall\ 0 \le t < t_{u_0},$$

for some $t_{u_0} > 0$, where $u_\lambda(t,u_0)$ is the solution of (2.1); and

3. *if $(x_\lambda(t), y_\lambda(t))$ is a solution of (2.10), then there are a $\beta_\lambda > 0$ and a $k_\lambda > 0$ with k_λ depending on $(x_\lambda(0), y_\lambda(0))$ such that*

$$\|y_\lambda(t) - h(x_\lambda(t),\lambda)\|_H \le k_\lambda e^{-\beta_\lambda t}.$$

If we consider only the existence of the local center manifold, then conditions in (2.9) can be modified in the following fashion. Let the operator $L_\lambda = \mathcal{L}_1^\lambda \oplus \mathcal{L}_2^\lambda$ and $\mathcal{L}_2^\lambda$ be decomposed into

$$\begin{cases} \mathcal{L}_2^\lambda = \mathcal{L}_{21}^\lambda \oplus \mathcal{L}_{22}^\lambda, \\ E_2^\lambda = E_{21}^\lambda \oplus E_{22}^\lambda,\ \widetilde{E}_2^\lambda = \widetilde{E}_{21} \oplus \widetilde{E}_{22}^\lambda, \\ \dim E_{21}^\lambda = \dim \widetilde{E}_{21}^\lambda < \infty, \\ \mathcal{L}_{2i}^\lambda : E_{2i}^\lambda \to \widetilde{E}_{2i}^\lambda \quad \text{are invariant } (i = 1,2), \end{cases} \tag{2.11}$$

such that at $\lambda = \lambda_0$

$$\begin{cases} \text{eigenvalues of } \mathcal{L}_1^\lambda : E_1^\lambda \to \widetilde{E}_1^\lambda \text{ have zero real parts,} \\ \text{eigenvalues of } \mathcal{L}_{21}^\lambda : E_{21}^\lambda \to \widetilde{E}_{21}^\lambda \text{ have positive real parts,} \\ \text{eigenvalues of } \mathcal{L}_{22}^\lambda : E_{22}^\lambda \to \widetilde{E}_{22}^\lambda \text{ have negative real parts.} \end{cases} \tag{2.12}$$

Then we have the following center manifold theorem.

Theorem 2.4. *Assume (2.4)–(2.6), (2.8), (2.11), and (2.12). Then the conclusions (1) and (2) in Theorem 2.3 hold true.*

2.3. Attractor bifurcation. A complex number $\beta = \alpha_1 + i\alpha_2 \in \mathbb{C}$ is called an eigenvalue of $L_\lambda : H_1 \longrightarrow H$ if there are $x, y \in H_1$ such that

$$L_\lambda z = \beta z, \quad z = x + iy,$$

or, equivalently,

$$\begin{aligned} L_\lambda x &= \alpha_1 x - \alpha_2 y, \\ L_\lambda y &= \alpha_2 x + \alpha_1 y. \end{aligned}$$

Now let the eigenvalues (counting the multiplicity) of L_λ be given by

$$\beta_1(\lambda), \beta_2(\lambda), \dots, \beta_k(\lambda) \in \mathbb{C}.$$

Suppose that

$$\operatorname{Re} \beta_i(\lambda) = \begin{cases} < 0 & \text{if } \lambda < \lambda_0, \\ = 0 & \text{if } \lambda = \lambda_0 \\ > 0 & \text{if } \lambda > \lambda_0, \end{cases} \qquad (1 \leq i \leq m+1), \tag{2.13}$$

$$\operatorname{Re} \beta_j(\lambda_0) < 0 \qquad \forall\, m+2 \leq j. \tag{2.14}$$

Let the eigenspace of L_λ at λ_0 be

$$E_0 = \bigcup_{i=1}^{m+1} \bigcup_{k=1}^{\infty} u \in H_1 | (L_{\lambda_0} - \beta_i(\lambda_0))^k u = 0.$$

It is known that $\dim E_0 = m + 1$.

Let $H_1 = H = \mathbb{R}^n$. The following attractor bifurcation theorem was proved in [9].

Theorem 2.5 (attractor bifurcation theorem). *Let $H_1 = H = \mathbb{R}^n$, the conditions (2.13) and (2.14) hold true, and $u = 0$ be a locally asymptotically stable equilibrium point of (2.1) at $\lambda = \lambda_0$. Then the following assertions hold true.*

1. *Equation (2.1) bifurcates from $(u, \lambda) = (0, \lambda_0)$ an attractor $\mathcal{A}_\lambda$ for $\lambda > \lambda_0$, with $m \leq \dim \mathcal{A}_\lambda \leq m + 1$, which is connected as $m > 0$.*
2. *The attractor $\mathcal{A}_\lambda$ is a limit of a sequence of $(m+1)$-dimensional annulus M_k with $M_{k+1} \subset M_k$. In particular, if $\mathcal{A}_\lambda$ is a finite simplicial complex, then $\mathcal{A}_\lambda$ has the homotopy type of S^m.*
3. *For any $u_\lambda \in \mathcal{A}_\lambda$, u_λ can be expressed as*

$$u_\lambda = v_\lambda + o(\|v_\lambda\|_{H_1}), \quad v_\lambda \in E_0.$$

4. *If $G : H_1 \to H$ is compact and the equilibrium points of (2.1) in $\mathcal{A}_\lambda$ are finite, then we have the index formula*

$$\sum_{u_i \in \mathcal{A}_\lambda} ind[-(L_\lambda + G), u_i] = \begin{cases} 2 & \text{if } m = \text{even}, \\ 0 & \text{if } m = \text{odd}. \end{cases}$$

5. *If $u = 0$ is globally stable for (2.1) at $\lambda = \lambda_0$, then for any bounded open set $U \subset H$ with $0 \in U$ there is an $\varepsilon > 0$ such that as $\lambda_0 < \lambda < \lambda_0 + \varepsilon$, the attractor $\mathcal{A}_\lambda$ bifurcated from $(0, \lambda_0)$ attracts U/Γ in H, where Γ is the stable manifold of $u = 0$ with codimension $m + 1$. In particular, if (2.1) has a global attractor in H, then $U = H$.*

Remark 2.6. As H_1 and H are infinite-dimensional Hilbert spaces, if (2.1) satisfies the conditions (2.4)–(2.6), (2.13), and (2.14) and $u = 0$ is a locally (global) asymptotically stable equilibrium point of (2.1) at $\lambda = \lambda_0$, then the assertions (1)–(5) of Theorem 2.5 hold; see [9, 8].

3. Bifurcation of the GL equation with Dirichlet boundary condition.

As mentioned in the introduction, we study in this article attractor bifurcation of the GL equation under either the Dirichlet or the periodic boundary conditions.

We start with the GL equation with the Dirichlet boundary condition. Let

$$H^k(\Omega, \mathbb{C}) = \{u_1 + iu_2 \mid u_j \in H^k(\Omega),\ j = 1, 2\},$$
$$H_0^1(\Omega, \mathbb{C}) = \{u \in H^1(\Omega, \mathbb{C}) \mid u|_{\partial\Omega} = 0\},$$

where $H^k(\Omega)$ is the usual real-valued Sobolev space.

Let λ_1 be the first eigenvalue of $-\triangle$ with the Dirichlet boundary condition (1.4). Then we have the following main bifurcation theorem for the GL equations with the Dirichlet boundary condition.

Theorem 3.1.

1. *If $\lambda \leq \alpha\lambda_1$, $u = 0$ is a globally asymptotically stable equilibrium point of (1.1) with (1.4).*
2. *As λ crosses $\alpha\lambda_1$, i.e., for any $\alpha\lambda_1 < \lambda < \alpha\lambda_1 + \epsilon$ for some $\epsilon > 0$, the problem (1.1) with (1.4) bifurcates from $(u, \lambda) = (0, \alpha\lambda_1)$ an attractor Σ_λ.*
3. *The bifurcated attractor Σ_λ has dimension between 1 and 2 and is a limit of a sequence of 2D annuli M_k with $M_{k+1} \subset M_k$; i.e., $\Sigma_\lambda = \cap_{k=1}^{\infty} M_k$.*
4. *If $\beta \neq 0$, then the bifurcation is a Hopf bifurcation, i.e., $\Sigma_\lambda = S^1$, which is asymptotically stable (limiting cycle).*
5. *If $\beta = 0$ and $\rho \neq 0$, then the bifurcated attractor Σ_λ is a periodic orbit, which is a limiting cycle.*
6. *Moreover, for each $\alpha\lambda_1 < \lambda < \alpha\lambda_1 + \epsilon$, the bifurcated attractor Σ_λ attracts the open set $L^2(\Omega, \mathbb{C})/\Gamma$, where Γ is the stable manifold of $u = 0$ having codimension two in $L^2(\Omega, \mathbb{C})$.*

Proof. We proceed in several steps as follows.

Step 1. Let $u = u_1 + iu_2$. The GL problem (1.1) with (1.3) can be equivalently written as follows:

$$\begin{cases} \dfrac{\partial u_1}{\partial t} = \alpha\triangle u_1 - \beta\triangle u_2 + \lambda u_1 - \sigma|u|^2 u_1 + \rho|u|^2 u_2, \\ \dfrac{\partial u_2}{\partial t} = \beta\triangle u_1 + \alpha\triangle u_2 + \lambda u_2 - \sigma|u|^2 u_2 - \rho|u|^2 u_1, \\ u_1(x, 0) = \phi(x), \quad u_2(x, 0) = \psi(x). \end{cases} \tag{3.1}$$

We shall apply Theorems 2.3 and 2.5 to prove this theorem. Let

$$H_1 = H^2(\Omega, \mathbb{C}) \cap H_0^1(\Omega, \mathbb{C}), \quad H = L^2(\Omega, \mathbb{C}).$$

The mappings $L_\lambda = -A + B_\lambda$ and $G : H_1 \to H$ are defined as

$$\begin{aligned} -Au &= \begin{pmatrix} \alpha\triangle u_1 - \beta\triangle u_2 \\ \beta\triangle u_1 + \alpha\triangle u_2 \end{pmatrix}, \\ B_\lambda u &= \lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \\ Gu &= \begin{pmatrix} -\sigma|u|^2 u_1 + \rho|u|^2 u_2 \\ -\sigma|u|^2 u_2 - \rho|u|^2 u_1 \end{pmatrix}. \end{aligned}$$

It is known that $H_{1/2} = H_0^1(\Omega, \mathbb{C})$. By the Sobolev embedding theorems and $1 \le n \le 3$, the mapping $G : H_{1/2} \to H$ is C^∞. The condition (2.4) is fulfilled.

Let $\{\lambda_k\} \subset \mathbb{R}$ and $\{e_k\} \subset H^2(\Omega) \cap H_0^1(\Omega)$ be the eigenvalues and eigenvectors of $-\triangle$ with the Dirichlet boundary condition (1.4)

$$\begin{cases} -\triangle e_k = \lambda_k e_k, \\ e_k|_{\partial\Omega} = 0. \end{cases}$$

We know that

$$0 < \lambda_1 < \lambda_2 \le \cdots \le \lambda_k \le \cdots, \quad \lambda_k \to \infty \text{ as } k \to \infty,$$

and $\{e_k\}$ is an orthogonal basis of $L^2(\Omega)$.

It is easy to see that the eigenvalues of A are given by

$$\alpha\lambda_k \pm i\beta\lambda_k, \quad k = 1, 2, \ldots,$$

with the corresponding eigenvectors

$$z_k = e_k + ie_k,$$

and $\{e_k, ie_j \mid 1 \le k, j < \infty\}$ is an orthogonal basis of H. Thus the conditions (2.5) and (2.6) are valid for A and B_λ. The eigenvalues of $L_\lambda = -A + B_\lambda$ are as follows:

$$(\lambda - \alpha\lambda_k) \pm i\beta\lambda_k, \qquad k = 1, 2, \ldots. \tag{3.2}$$

In addition, the spaces H and H_1 can be decomposed into the form

$$\begin{aligned} &H_1 = E_1 \oplus E_2 \quad \text{and} \quad H = E_1 \oplus \widetilde{E}_2, \\ &E_1 = \{x_1 e_1 + iy_1 e_1 \mid x_1, y_1 \in \mathbb{R}\}, \\ &E_2 = \left\{ \sum_{k=2}^\infty (x_k + iy_k) e_k \;\middle|\; \sum_{k=2}^\infty \lambda_k^2 (x_k^2 + y_k^2) < \infty \right\}, \\ &\widetilde{E}_2 = \left\{ \sum_{k=2}^\infty (x_k + iy_k) e_k \;\middle|\; \sum_{k=2}^\infty (x_k^2 + y_k^2) < \infty \right\}, \end{aligned}$$

and the operator L_λ is decomposed into

$$\begin{cases} L_\lambda = \mathcal{L}_1^\lambda \oplus \mathcal{L}_2^\lambda, \\ \mathcal{L}_1^\lambda = L_\lambda|_{E_1} : E_1 \to E_1, \quad \mathcal{L}_2^\lambda = L_\lambda|_{E_2} : E_2 \to \widetilde{E}_2. \end{cases}$$

Thus the conditions (2.8) and (2.9) are satisfied.

By the center manifold theorem, the attractor bifurcation of (1.1) with (1.4) is equivalent to that of the bifurcation equations

$$\begin{cases} \dfrac{dx}{dt} = (\lambda - \alpha\lambda_1)x_1 + \beta\lambda_1 y_1 + P_1G_1(x_1 + iy_1 + h), \\ \dfrac{dy}{dt} = -\beta\lambda_1 x_1 + (\lambda - \alpha\lambda_1)y_1 + P_1G_2(x_1 + iy_1 + h), \end{cases} \tag{3.3}$$

where $h = h_1 + ih_2$ is the center manifold function satisfying

$$h(x_1, y_1) = o(|x_1| + |y_1|),$$

and $P_1G_i(u)$ $(i = 1, 2)$ are given by

$$\begin{aligned} &P_1G_1(u) = \int_\Omega [-\sigma|u|^2u_1 + \rho|u|^2u_2]e_1 dx, \\ &P_1G_2(u) = \int_\Omega [-\sigma|u|^2u_2 - \rho|u|^2u_1]e_1 dx, \\ &u = u_1 + iu_2 = \sum_{k=1}^\infty (x_k + iy_k)e_k. \end{aligned} \tag{3.4}$$

Step 2. Now we show that $u = 0$ is a globally asymptotically stable equilibrium of (2.1) for $\lambda \leq \alpha\lambda_1$. In fact, from (3.1) we can derive that

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_\Omega |u|^2 dx &= \int_\Omega (-\alpha|\nabla u|^2 + \lambda|u|^2 - \sigma|u|^4)dx \\ &\leq -\int_\Omega [(\alpha\lambda_1 - \lambda)|u|^2 + \sigma|u|^4]dx \end{aligned}$$

which implies that $u = 0$ is globally stable.

Step 3. We know that (1.1) has a global attractor; see [12]. Obviously, for the eigenvalues (3.2) of L_λ, the conditions (2.13) and (2.14) for $\lambda_0 = \alpha\lambda_1$ are satisfied. Therefore, by Remark 2.6, (2.1) bifurcates from $(u, \lambda) = (0, \alpha\lambda_1)$ an attractor Σ_λ which attracts H/Γ.

Step 4. We now prove that $\Sigma_\lambda = S^1$.

Obviously, as $\beta \neq 0$, the bifurcation is the typical Hopf bifurcation. Therefore, we have to consider only the case where $\beta = 0$. In this case, the bifurcation equations (3.3) read

$$\begin{cases} \dfrac{dx}{dt} = \varepsilon x + P_1G_1(xe_1 + h_1 + iye_1 + ih_2), \\ \dfrac{dy}{dt} = \varepsilon y + P_1G_2(xe_1 + h_1 + iye_1 + ih_2), \end{cases} \tag{3.5}$$

where $\varepsilon = \lambda - \alpha\lambda_1 > 0$ sufficiently small. By (3.4) we have

$$\begin{aligned} P_1G_1(x,y) &= \int_\Omega [-\sigma u_1^3 + \rho u_2^3 - \sigma u_2^2 u_1 + \rho u_1^2 u_2] e_1 dx \\ &= (\text{by } h(x,y) = o(|x| + |y|)) \\ &= a(-\sigma x^3 + \rho y^3 - \sigma y^2 x + \rho x^2 y) + o(|x|^3 + |y|^3), \\ P_1G_2(x,y) &= a(-\sigma y^3 - \rho x^3 - \sigma x^2 y - \rho y^2 x) + o(|x|^3 + |y|^3), \end{aligned}$$

where $u_1 = xe_1 + h_1(x,y)$, $u_2 = ye_1 + h_2(x,y)$, and

$$a = \int_\Omega e_1^4(x) dx > 0.$$

Thus, the bifurcation equations (3.5) lead to

$$\text{(3.6)} \qquad \begin{cases} \dfrac{dx}{dt} = \varepsilon x - a(\sigma x^3 - \rho y^3 + \sigma y^2 x - \rho x^2 y) + o(|x|^3 + |y|^3), \\ \dfrac{dy}{dt} = \varepsilon y - a(\sigma y^3 + \rho x^3 + \sigma x^2 y + \rho y^2 x) + o(|x|^3 + |y|^3). \end{cases}$$

We can see that the attractor Σ_λ has no nonzero singular point; i.e., the singular point $u = 0$ of (1.1) with (1.4) is unique provided $|\rho| + |\beta| \neq 0$, because from (3.1) we have

$$\int_\Omega \left[u_2 \frac{\partial u_1}{\partial t} - u_1 \frac{\partial u_2}{\partial t} \right] dx = \int_\Omega [\beta |\nabla u|^2 + \rho |u|^4] dx.$$

By Theorem 2.5, Σ_λ has the homotopy type of S^1; hence Σ_λ contains at least one periodic orbit provided $\rho \neq 0$.

Take the polar coordinate system

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Then (3.6) becomes

$$\text{(3.7)} \qquad \begin{cases} \dfrac{dr}{d\theta} = \dfrac{\varepsilon - a\sigma r^2 + o(r^2)}{a\rho r}, \\ r(0) = r_0. \end{cases}$$

From (3.7) it follows that

$$\frac{a\rho}{2}(r^2(2\pi) - r^2(0)) = \int_0^{2\pi} [\varepsilon - a\sigma r^2 + o(r^2)] d\theta.$$

Because $r^2 = r^2(\theta, r_0)$ is C^∞ on $r_0 \geq 0$, we have the Taylor expansion

$$r^2(\theta, r_0) = r_0^2 + R(\theta) \cdot o(|r_0|^2), \quad R(0) = 0.$$

Hence we get

$$\frac{a\rho}{2}(r^2(2\pi) - r^2(0)) = 2\pi\varepsilon - 2\pi a\sigma r_0^2 + o(|r_0|^2).$$

Obviously the initial values $r_0 > 0$ in (3.7) satisfying

$$2\pi\varepsilon - 2\pi a\sigma r_0^2 + o(|r_0|^2) = 0 \tag{3.8}$$

are corresponding to the periodic orbits of (3.6). It is easy to see that the solution $r_0^2 > 0$ of (3.8) near $r_0 = 0$ is unique. Thus we deduce that Σ_λ is a periodic orbit provided $\rho \neq 0$.

The proof is complete. ■

4. Bifurcation of the GL equation with periodic boundary condition.

For the GL equation with periodic boundary condition, the first eigenspace is larger than that in the Dirichlet boundary condition case, and to proceed, we need the following function spaces:

$$H^k_{per}(\Omega, \mathbb{C}) = \{u \in H^k(\Omega, \mathbb{C}) \mid u \text{ satisfy } (1.5)\}.$$

Then the main result in this section is the following theorem.

Theorem 4.1. *For the GL equation (1.1) with the periodic boundary condition (1.5), we have the following assertions.*

1. (a) *As $\lambda > \alpha$, the problem (1.1) with (1.5) bifurcates from $(u, \lambda) = (0, \alpha)$ an invariant state Σ_λ. The bifurcated attractor Σ_λ has dimension between $4n - 1$ and $4n$ and is a limit of a sequence of $4n$ annulus M_k with $M_{k+1} \subset M_k$; i.e., $\Sigma_\lambda = \cap_{k=1}^\infty M_k$.*
   (b) *If $|\rho| + |\beta| \neq 0$, then Σ_λ contains no steady state solutions of (1.1) with (1.5).*
2. (a) *As $\lambda \leq 0$, $u = 0$ is globally asymptotically stable.*
   (b) *As $\lambda > 0$, the problem (1.1) with (1.5) bifurcates from $(u, \lambda) = (0, 0)$ an attractor $\Sigma_\lambda \subset L^2(\Omega, \mathbb{C})$. The bifurcated attractor Σ_λ has dimension between 1 and 2 and is a limit of a sequence of 2D annuli M_k with $M_{k+1} \subset M_k$; i.e., $\Sigma_\lambda = \cap_{k=1}^\infty M_k$.*
   (c) *If $\rho \neq 0$, then Σ_λ is a periodic orbit.*
   (d) *Σ_λ attracts $L^2(\Omega, \mathbb{C})/\Gamma$, where Γ is the stable manifold of $u = 0$ with codimension two in $L^2(\Omega, \mathbb{C})$.*

Proof. Let $H_1 = H^2_{per}(\Omega, \mathbb{C})$, $H = L^2_{per}(\Omega, \mathbb{C})$, and the mappings L_λ and $G : H_1 \to H$ be as defined in the previous section. Similar to the proof of Theorem 3.1, L_λ and G satisfy the conditions in Theorem 2.3.

We know that the eigenvalue problem

$$\begin{cases} -\triangle e_k = \lambda_k e_k, \\ e_k(x + 2k\pi) = e_k(x) \end{cases} \tag{4.1}$$

has an eigenvalue sequence

$$\lambda_0 = 0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots, \quad \lambda_k \to \infty \text{ as } k \to \infty,$$

and an eigenvector sequence $\{e_k\}$ which constitutes a common orthogonal basis of H_1 and H. The second eigenvalue $\lambda_1 = 1$ has multiplicity $2n$, i.e., $\lambda_1 = \cdots = \lambda_{2n}$, with the first eigenvectors

$$\sin x_j, \quad \cos x_j \qquad (x = (x_1, \dots, x_n) \in \Omega = [0, 2\pi]^n).$$

Eigenvalues of L_λ are as in (3.2), and the second eigenvalue $\Lambda_1 = (\lambda - \alpha) \pm i\beta$ has multiplicity $4n$. For simplicity, let

$$e_{2j-1} = \sin x_j, \quad e_{2j} = \cos x_j \qquad (j = 1, \dots, n).$$

Then the spaces H and H_1 can be decomposed into the following form:

$$\begin{aligned} &H = E_1 \oplus \widetilde{E}_2, \\ &E_1 = \left\{ \sum_{j=1}^{2n} (z_{1j} + iz_{2j}) e_j \;\middle|\; z_{1j}, z_{2j} \in \mathbb{R} \right\}, \\ &\widetilde{E}_2 = E_1^\perp. \end{aligned}$$

Then the bifurcation equations of (1.1) with (1.5) are given by

$$\left\{ \begin{aligned} \frac{dZ_1}{dt} &= (\lambda - \alpha) Z_1 + \beta Z_2 + PG_1(u), \\ \frac{dZ_2}{dt} &= -\beta Z_1 + (\lambda - \alpha) Z_2 + PG_2(u), \end{aligned} \right. \tag{4.2}$$

where $u = u_1 + iu_2$ and

$$\begin{aligned} &(u_1, u_2) = (Z_1 + h_1(Z_1, Z_2), Z_2 + h_2(Z_1, Z_2)), \\ &(Z_1, Z_2) = \sum_{j=1}^{2n} (z_{1j}, z_{2j}) e_j. \end{aligned}$$

Here $h = h_1 + ih_2 : E_1 \to \widetilde{E}_2$ is the center manifold function satisfying

$$h(Z_1, Z_2) = o(|Z_1| + |Z_2|) \tag{4.3}$$

and

$$\begin{aligned} PG_1(u) &= \sum_{j=1}^{2n} e_j \int_\Omega [-\sigma |u|^2 u_1 + \rho |u|^2 u_2] e_j dx, \\ PG_2(u) &= \sum_{j=1}^{2n} e_j \int_\Omega [-\sigma |u|^2 u_1 - \rho |u|^2 u_2] e_j dx. \end{aligned}$$

By Theorem 2.5, we infer from (4.3) that (1.1) and (1.5) bifurcate from $(u, \lambda) = (0, \alpha)$ an invariant set Σ_λ.

The proof is complete. ■

Remark 4.2. In fact, the invariant set Σ_λ of (1.1) with (1.5) is a sphere S^{4n-1}; namely, Σ_λ is homeomorphic to a sphere S^{4n-1}. The topological structure of an attractor of vector fields should be stable provided some nondegenerate conditions hold. We shall discuss this topic elsewhere.

5. Bifurcation of invariant sphere S^m. More generally, for the GL equations we have the bifurcation theorem of invariant sphere S^m ($m \geq 1$) at any eigenvalue.

Theorem 5.1. *Let λ_k be an eigenvalue of $-\triangle$ with the boundary condition (1.4), or (1.5), which has multiplicity $m \geq 1$. Then, as $\lambda > \alpha\lambda_k$, the problem (1.1) with (1.4), or (1.1) with (1.5), bifurcates from $(u,\lambda) = (0, \alpha\lambda_k)$ an invariant set Σ_λ. This invariant set Σ_λ has dimension between $2m-1$ and $2m$ and is a limit of a sequence of $2m$-dimensional annuli M_k with $M_{k+1} \subset M_k$; i.e., $\Sigma_\lambda = \cap_{k=1}^{\infty} M_k$. If $|\beta| + |\rho| \neq 0$, then there is no singular point in Σ_λ.*

Proof. We denote the eigenvectors of $-\triangle$ corresponding to λ_k by

$$\{e_1^*, \ldots, e_m^*\}.$$

Thus, the space H_1 and H defined in Theorem 3.1 or Theorem 4.1 can be decomposed into

$$\begin{aligned}
&H_1 = E_m \oplus E_m^\perp, \quad H = \widetilde{E}_m \oplus \widetilde{E}_m^\perp,\\
&E_m = \operatorname{span}\{e_i^* + ie_j^* \mid 1 \leq i, j \leq m\},\\
&E_m^\perp = \{u \in H_1 \mid \langle u, v\rangle_{H_1} = 0 \ \forall v \in E_m\},\\
&\widetilde{E}_m = E_m, \quad \widetilde{E}_m^\perp = \{u \in H \mid \langle u, v\rangle_H = 0 \ \forall v \in \widetilde{E}_m\}.
\end{aligned}$$

By the center manifold theorem, the bifurcation equations of (1.1) with (1.4), or (1.1) with (1.5), at $\lambda = \lambda_k$ are equivalent to

$$\begin{cases}
\dfrac{\partial v_1}{\partial t} = \alpha\triangle v_1 - \beta\triangle v_2 + \lambda v_1 + PG_1(v + h(v)),\\
\dfrac{\partial v_2}{\partial t} = \beta\triangle v_1 + \alpha\triangle v_2 + \lambda v_2 + PG_2(v + h(v)),
\end{cases} \tag{5.1}$$

where λ is near λ_k, $v = v_1 + iv_2 \in E_m$, and $h : E_m \to E_m^\perp$ is the center manifold function, $G = (G_1, G_2) : H_1 \to H$ defined as in Theorem 3.1 or Theorem 4.1, and $P : H \to \widetilde{E}_m$ is the projection.

The equations (5.1) are a system of ordinary differential equations with order $2m$:

$$\begin{cases}
\dfrac{dZ_1}{dt} = (\lambda - \alpha)Z_1 + \beta Z_2 + [-\sigma|Z|^2 Z_1 + \rho|Z|^2 Z_2] + o(|Z|^3),\\
\dfrac{dZ_2}{dt} = -\beta Z_1 + (\lambda - \alpha)Z_2 + [-\sigma|Z|^2 Z_2 - \rho|Z|^2 Z_1] + o(|Z|^3),
\end{cases} \tag{5.2}$$

where $Z = Z_1 + iZ_2$. The eigenvalues of the linear part are still $(\lambda - \alpha\lambda_k) \pm i\beta\lambda_k$, with multiplicity $2m$.

By Theorem 2.5 it suffices to prove that $v = 0$ is asymptotically stable for (5.2) at $\lambda = \alpha\lambda_k$.

For $\lambda = \alpha\lambda_k$, we infer from (5.2) that

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\int_\Omega |v|^2 dx &= \int_\Omega G(v+h(v))v dx\\ &= (\text{by } h(v) = o(|v|))\\ &= \int_\Omega G(v)v dx + o(|v|^4)\\ &= -\sigma \int_\Omega |v|^4 dx + o(|v|^4),\end{aligned}$$

which implies that $v = 0$ is asymptotically stable for (5.2) at $\lambda = \alpha\lambda_k$. The proof is complete. ■

Acknowledgments. The authors thank the two anonymous referees for their insightful comments.

REFERENCES

[1] M. Bartuccelli, P. Constantin, C. R. Doering, J. D. Gibbon, and M. Gisselfält, *Hard turbulence in a finite-dimensional dynamical system?*, Phys. Lett. A, 142 (1989), pp. 349–356.

[2] M. Bartuccelli, P. Constantin, C. R. Doering, J. D. Gibbon, and M. Gisselfält, *On the possibility of soft and hard turbulence in the complex Ginzburg-Landau equation*, Phys. D, 44 (1990), pp. 421–444.

[3] C. R. Doering, J. D. Gibbon, D. D. Holm, and B. Nicolaenko, *Low-dimensional behaviour in the complex Ginzburg-Landau equation*, Nonlinearity, 1 (1988), pp. 279–309.

[4] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[5] H. G. Kaper and P. Takáč, *Ginzburg-Landau dynamics with a time-dependent magnetic field*, Nonlinearity, 11 (1998), pp. 291–305.

[6] I. Kukavica, *An upper bound for the winding number for solutions of the Ginzburg-Landau equation*, Indiana Univ. Math. J., 41 (1992), pp. 825–836.

[7] I. Kukavica, *Hausdorff length of level sets for solutions of the Ginzburg-Landau equation*, Nonlinearity, 8 (1995), pp. 113–129.

[8] T. Ma and S. Wang, *Attractor bifurcation theory and its applications to Rayleigh-Bénard convection*, Comm. Pure Appl. Anal., 2 (2003), pp. 591–599.

[9] T. Ma and S. Wang, *Dynamic bifurcation of nonlinear evolution equations*, Chinese Ann. Math. Ser. B, to appear.

[10] T. Ma and S. Wang, *Bifurcation Theory and Applications*, World Scientific, River Edge, NJ, 2005, to appear.

[11] Q. Tang and S. Wang, *Time dependent Ginzburg-Landau equations of superconductivity*, Phys. D, 88 (1995), pp. 139–166.

[12] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Appl. Math. Sci. 68, Springer-Verlag, New York, 1997.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 636–670

Minimal Models of Bursting Neurons: How Multiple Currents, Conductances, and Timescales Affect Bifurcation Diagrams[*]

R. M. Ghigliazza[†] and P. Holmes[‡]

Abstract. After reviewing the Hodgkin–Huxley ionic current formulation, we introduce a three-variable generic model of a single-compartment neuron comprising a two-dimensional fast subsystem and a very slow recovery variable. We study the effects of fast and slow currents on the existence and stability of equilibria and periodic orbits for the fast subsystem, presenting a classification of currents and developing graphical tools that aid in the analysis and construction of models with specified properties. We draw on these to propose a minimal model of a bursting neuron, identifying biophysical parameters that can shape and regulate key characteristics of the membrane voltage pattern: bursting frequency, duty cycle, spike rate, and the number of action potentials per burst. We present additional examples from the literature for comparison and illustration, and in a companion paper [*SIAM J. Appl. Dyn. Syst.*, 3 (2004), pp. 671–700], we construct a model of an insect central pattern generator using these methods.

Key words. bursting neurons, motoneurons, fast-slow systems, bifurcation, stability

AMS subject classifications. 34C15, 34C25, 34C29, 37Gxx, 92B05, 92B20, 92C20

DOI. 10.1137/030602307

1. Introduction.

In this and a companion paper [1] we develop and analyze a generic model of a bursting neuron and assemble a set of such models, suitably adapted to interneurons and motoneurons, to model a central pattern generator (CPG) for insect locomotion. We have two main goals: to integrate and extend a body of work, largely in theoretical and mathematical neuroscience, that enables (semi-) analytical studies of bursting neurons, while maintaining sufficient biophysical detail for comparisons with experimental data; and to use this to derive a model of a CPG that reveals how key locomotive properties may be determined by individual neurons and the network as a whole. In this first paper we show how complex models can be reduced and develop the analytical methods; in [1] we construct the CPG model.

The first dynamical neural model based on biophysical data was due to Hodgkin and Huxley [2], and their description of the action potential (AP) and ionic currents in the giant axon of the squid has been vastly extended and generalized in the half century since. Detailed axonal and dendritic geometry can be included, for example, at the unicellular level. However,

[*] Received by the editors December 8, 2003; accepted for publication (in revised form) by M. Golubitsky July 2, 2004; published electronically December 27, 2004. This work was supported by DoE: DE-FG02-95ER25238.

http://www.siam.org/journals/siads/3-4/60230.html

[†] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (rghiglia@princeton.edu). This author was partially supported by a Burroughs-Wellcome Training Grant in Biological Dynamics: 1001782, and a Britt and Eli Harari Fellowship from the Department of Mechanical and Aerospace Engineering, Princeton University.

[‡] Department of Mechanical and Aerospace Engineering and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (pholmes@math.princeton.edu).

behaviors even as simple as scratching or breathing require *networks* of neurons, and the resulting firing patterns depend on three levels of activity: intracellular, synaptic, and network. Models ignoring any of these levels risk oversimplification [3], perhaps especially in invertebrates, in which relatively few neurons may be responsible for such diverse behaviors as searching, walking, and running [4, 5]. The wealth of neurophysiological data collected since Hodgkin and Huxley's paper has led to rather complicated models (e.g., [6, 7, 8]), some multicompartmental and including seven or more ionic currents, that require on the order of ten ODEs and fifty parameters per cell. These models are specific to particular animals and even to in vitro preparations and, not being amenable to analytical comparative studies, do not readily reveal general principles. In this paper and [1] we seek a balance between such complexity and simpler phenomenological models employing phase oscillators [9, 10] or connectionist circuits [11] that have been used to study network connectivity effects.

Bursting oscillations have been widely studied, mostly at the single-cell level, e.g., [12, 13, 14, 15, 16, 17]; general classifications have been proposed [18, 19, 20, 21, 22], and polynomial reductions have been developed and thoroughly analyzed [23, 24, 25]. Some network studies have also recently been done [17, 26]. When limited experimental data is available, as in [17] and the CPG model of [1], generic models and broad parameter variations can still lead to testable hypotheses and provide motivation to verify novel predictions [27]. However, while asymptotic reductions and polynomial approximations aid mathematical analyses, they often obscure biophysical effects that must be retained if one is to understand how internal components and architecture, as well as proprioceptive sensing and commands from higher centers, can influence a network [28].

Single-current effects on individual cells are qualitatively understood, but collective influences have not been fully explored. We therefore devote this first paper to analyzing how multiple (fast and slow) currents conspire to affect the location and stability of equilibria and limit cycles, with a view to determining how biophysical parameters can effect changes in behavioral variables such as spike rate, bursting frequency, duty cycle, and number of APs per burst. Our methods identify currents which are unessential to the bursting mechanism, suggest dimensional reductions, and provide guidelines for "designing" bursters with desired behaviors when intracellular biophysical data is lacking. Thus, while they are used in [1] to model an insect CPG, these methods offer a more general set of tools for studying the neural basis of rhythmogensis. Indeed, comparisons with several existing models are noted in passing, and a specific example is given in section 4.3. In developing these tools, we have profited from many earlier studies, including those of Rose and Hindmarsh [29, 14, 16] (on $I-v$ steady state curves) and of FitzHugh [30] and Rose and Hindmarsh [14, 31] (on combining gating variables). Here we treat only third order models with one very slow recovery variable; we note that Smolen, Terman, and Rinzel [32] considered two slow variables. Specific references to these and other relevant papers and models will be made in the course of the paper.

This paper is organized as follows. After reviewing basic ideas of single-compartment ion-channel models and noting the role of disparate timescales in section 2, we describe a three-variable generic model of a bursting neuron in section 3. We identify and analyze the effects of individual current and conductance parameters on branches of equilibria and periodic orbits and their bifurcations, and then in section 4 we lay out a "minimal" model for a bursting neuron, of sufficient flexibility to represent both interneurons and motoneurons in the CPG application of [1]. We identify biophysical parameters that shape the bursting pattern and hence will determine key behavioral characteristics, and we illustrate further by showing how an example (the Sherman–Rinzel–Keizer (SRK) model [13]) can be modified to produce different behaviors. We summarize in section 5.

2. Ion-channel models.

Bursting, the clustering of spikes followed by a refractory period of relative quiescence, can vary substantially in form and function [22, 33, 21]. The mechanism can be described qualitatively as the interaction of two subsystems dynamically separated by their intrinsic time scales: a faster one, typically governed by sodium and potassium channels, which can either be at rest or exhibit (periodic) oscillations, and a slow subsystem driving the first through its quiescent and oscillatory states in a quasi-static manner [34, 35]. The slower mechanism can be attributed to the accumulation of intracellular calcium ions (referred to as calcium dynamics [33]) or to other slow voltage-dependent processes (e.g., [17]). In many cases, bursting models can therefore be framed as singularly perturbed systems [36]:

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}, c), \tag{2.1a}$$

$$\dot{c} = \delta g(\mathbf{u}, c), \tag{2.1b}$$

where the vector $\mathbf{u} = [v, \mathbf{w}] \in \mathbb{R}^{N+1}$, v denotes the cell membrane voltage, $\mathbf{w} = [w_1, \dots, w_N]^T$ represents a collection of N gating variables w_i to be explained below, and $\delta \ll 1$ is a small parameter. The variable c may represent calcium concentration or, more generally, any (very) slowly varying quantity responsible for bursting.

The subset of *fast* equations (2.1a) generally takes the Hodgkin–Huxley (HH) form [2] and can be written as follows [37]:

$$C\dot{v} = -I_{\text{ion}}(v, w_1, \dots, w_n, c) + I_{\text{ext}}(t), \tag{2.2a}$$

$$\dot{w}_i = \epsilon_i \frac{w_{i_\infty}(v) - w_i}{\tau_i(v)}, \qquad i = 1, \dots, N. \tag{2.2b}$$

The first equation (2.2a) describes the voltage dynamics, with C denoting the cell membrane capacitance, I_{ion} transmembrane ionic currents, and $I_{\text{ext}}(t)$ exogenous input currents, including synaptic and external inputs. Equations (2.2b) describe the first order kinetics of variables w_i that gate the ionic currents (see below), with ϵ_i a positive temperature-like parameter (not necessarily small). At steady state, gating variables approach voltage-dependent limits $w_{i_\infty}(v)$, usually described by sigmoidal functions:

$$w_{i_\infty}(v; k_{i_0}, v_{i_{th}}) = \frac{1}{1 + e^{-k_{i_0}(v - v_{i_{th}})}}, \tag{2.3}$$

where k_{i_0} determines the steepness of the transition occurring at a threshold potential $v_{i_{th}}$. Gating variables can be either *activating* ($k_{i_0} > 0$), with $w_{i_\infty} \approx 1$ for depolarized voltages $v > v_{i_{th}}$ and $w_{i_\infty} \approx 0$ for hyperpolarized levels $v < v_{i_{th}}$, or *inactivating* ($k_{i_0} < 0$), with $w_{i_\infty} \approx 1$ when hyperpolarized and $w_{i_\infty} \approx 0$ when depolarized.[1] The voltage-dependent "time constant" τ_i is generally described by

$$\tau_i(v; k_{i_0}, v_{i_{th}}) = \operatorname{sech}\left(k_{i_0}(v - v_{i_{th}})\right), \tag{2.4}$$

[1] It is sometimes useful to retain $k_{i_0} > 0$ in (2.3) and express inactivation via $1 - w_i$.

and, as implied in (2.3)–(2.4), the constants $k_{i_0}, v_{i_{th}}$ determining w_{i_∞} and τ_i are often taken to be the same for a given ion channel [38].

The term I_{ion} in (2.2a) is the sum of all ionic currents I_α. Ions move across the membrane via channels which are permeable to specific species (possibly more than one per channel), and they can be thought of as being in either of two states: open or closed. The total conductance associated with a given (sufficiently large) population of channels can be expressed as the (constant) maximal conductance $\bar{g}_\alpha$ for all channels open, multiplied by the fraction of open channels. Thus, each ionic current can generally be described as Ohmic and written in the form

$$I_\alpha(v, \mathbf{w}, c) = \bar{g}_\alpha \cdot \gamma_\alpha(v, w_1, \ldots, w_N, c) \cdot (v - E_\alpha); \tag{2.5}$$

more complicated "rectifying" conductances, such as those expressed by the Goldman–Hodgkin–Katz formula [39] can also be represented in this manner. Here E_α is the (Nernstian) reversal potential, α denotes the ion type, typically $\alpha \in \{\text{Na}, \text{K}, \text{Ca}, \text{Cl}, \text{L}\}$, L denoting the leakage current, and $\gamma_\alpha(v, \mathbf{w}, c)$ is a voltage-, gate-, and possibly c-dependent conductance factor for channels selective to ion α. To describe this dependence, Hodgkin and Huxley [2] introduced fictive gating particles and represented γ_α with one or two[2] activating and inactivating gating variables $w_i, w_j \in [0, 1]$, raised to integral powers a and b:

$$\gamma_\alpha(v, \mathbf{w}, c) = \zeta(v)\ \xi(c)\ w_i^a w_j^b. \tag{2.6}$$

The exponents a, b can be thought of as representing the number of subunits within a single channel necessary to open it; see Figure 1b. Probabilistic models based on this approach closely reproduce experimental data for large channel numbers [41].

A first possible simplification is to restrict the exponents a, b in (2.6) to unity. A rigorous approach would require a change of variable $z = w_i^a$, etc., as in [14], but two observations are pertinent. (i) Some models do have currents with exponent 1: e.g., Morris and Lécar [38] and extensions thereof to bursting models [37], I_{T} in Plant [42], I_{K} in Sherman, Rinzel, and Keizer [13], I_{K} in Keizer and Smolen [43], and $I_{\text{K(M)}}, I_{\text{K(C)}}, I_{\text{K(AHP)}}$ in [39, pp. 200–203]. (ii) More importantly, this restriction is not as severe as it may seem; for the steady state expression (2.3) at least, one can show that $w_\infty^a(v; k_0, v_{th})$ can be approximated by another sigmoidal function raised to the power 1 but with different coefficients $\bar{w}_\infty(v; \bar{k}_0, \bar{v}_{th})$. Taylor-expanding w_0^a, we can locally match the two functions to first order via the parameters $\bar{k}_0, \bar{v}_{th}$, and we have checked that for sigmoidal functions the pointwise match is acceptable, with maximum error of around 5% on the whole real line (results not shown).

In spite of the variety of ions and gating mechanisms, conductances come in two forms [44, 3]: persistent and transitory (see Figure 1). The names refer to steady state properties of w_{i_∞}: persistent activating or inactivating conductances being active, respectively, above or below a threshold, and transitory conductances being active only in a "window" of voltages. The former are described by a single gating variable, whereas a combination of activating and inactivating gating variables is used for transitory conductances. We will comment further on the functions $\zeta(v)$ and $\xi(c)$ in (2.6), but we anticipate that they can capture rectifying properties as described by the Goldman–Hodgkin–Katz equation [39], or "mixed" conductances, to be defined subsequently.

[2] An exception with three gating variables appeared in the model of Beeler and Reuter [40].

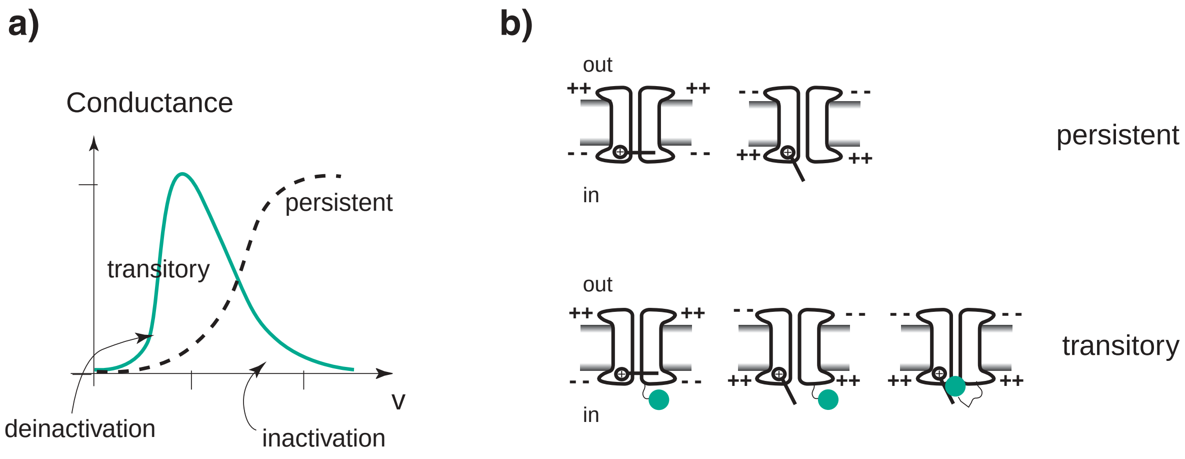


Figure 1. *Conductances in the HH formalism.* (a) *Qualitative dependence of persistent (dashed) and transitory conductances (solid).* (b) *Caricature of mechanisms underlying the opening and closure of the channels, with voltage increases to the right. Persistent conductances are represented by a voltage-dependent gating variable which continually opens the channel, allowing ionic transport; transitory conductances have an additional secondary mechanism that blocks the channel at voltages above the active range.*

We note that the coupling in (2.2) occurs only in the first equation (2.2a); gating variables are not directly coupled. This structure enables simplified analyses, as we now indicate (cf. [16]). A neuron may possess a dozen distinct ion channels [3], but if qualitative or semi-quantitative characteristics are adequate, a reduced model having fewer variables may suffice [30, 37]. If some ionic timescales τ_j in (2.2b) are significantly faster than others, we may (formally) set the corresponding gating variables at their equilibrium values $w_j = w_{j_\infty}(v)$. Likewise, functionally related variables with similar timescales may be lumped together (cf. [37]), and this is not atypical; see the comment below on hypothesis H1. Those variables whose channel dynamics have been equilibrated will henceforth be denoted by $n_i(v) = n_{i_\infty}(v)$; the (slower) gating variables w_j whose evolution equations are retained will be denoted by m_j. In general several such variables may be retained, but we shall henceforth restrict our attention to the case of a single slow variable, m, noting that in some cases this might represent a combination of two or more gating variables that move in step; see [30, 33] and the Rose–Hindmarsh model [14] in Appendix A.

The above reduction process, which was pioneered in FitzHugh's polynomial reduction of the HH model [30] (cf. [23, 14, 24]) and may be justified via geometric singular perturbation theory [36], considerably simplifies analyses but at the expense of obscuring some of the biophysics. We will therefore develop a three-dimensional model in this spirit but retaining the link to biophysical parameters. Pernarowski [24] and more recently De Vries [25] have amply demonstrated the richness and relevance of three-dimensional models in describing several distinct bursting behaviors.

3. A third order model.

Most neurons have many more membrane conductances than the two measured by Hodgkin and Huxley (e.g., Connor and Stevens [45], Plant [42], Chay and Keizer [12], McCormick and Huguenard [46]); one or two sodium conductances, two or three different types of calcium conductances, and many different potassium conductances are common [39]. In this section we will develop a framework to analyze the collective contributions of single ionic currents, with the goal of showing how biophysical parameters influence the existence and stability of equilibria and periodic orbits in the fast subsystem (2.2), and hence illuminating the global dynamics of the coupled fast-slow system (2.1).

We consider a class of models characterized by the following hypotheses:

H1. Existence of a single relatively slow (nonequilibrated) variable m in the fast subsystem.

H2. Multiplicative dependence of conductances on gating variables, voltage, and the very slow variable c: $\gamma(v, \mathbf{w}, c) = \zeta(v)\ \xi(c)\ \prod_i \sigma_{\alpha_i}(w_i)$.

The first hypothesis could be rephrased as "homogeneous dependence on one slow variable." In fact m may describe more than one channel, pairs of the form $\bar{g}_{\alpha_1}\sigma_{\alpha_1}(m)\zeta_{\alpha_1}(v)(v - E_{\alpha_1}) + \bar{g}_{\alpha_1}\sigma_{\alpha_2}(m)\zeta_{\alpha_2}(v)(v - E_{\alpha_2})$ being allowed. This is not as atypical as it may seem; see, e.g., the reduced models of Rose and Hindmarsh [14] (Appendix A) and Butera, Rinzel, and Smith [17] (Appendix C). The first hypothesis also implies that reduction to a three-dimensional system is possible (cf. [23, 22, 37, 25, 24] and references therein), and it allows for a wealth of different behaviors [35]. The second hypothesis formalizes a common assumption which holds for all models of which we are aware.

Under these hypotheses, we can formulate a rather general model. We will not commit to a particular choice of ionic currents until later, when we will be able to justify our choices for a "minimal" model. The principle channels allow the passage of four ions: sodium, potassium, chloride (or sometimes generic leakage ions), and calcium [3]. They are often highly selective, each admitting only one ionic specie. Calling the *relatively* slow gating variable m and using the above four ions, we can express the model as

$$C\dot{v} = -\sum_i I_{\alpha_i} + I_{\text{ext}}, \tag{3.1a}$$

$$\dot{m} = \frac{\epsilon}{\tau_m(v)}\left[m_\infty(v) - m\right], \tag{3.1b}$$

$$\dot{c} = \frac{\delta}{\tau_c(v)}\left[c_\infty(v) - c\right], \tag{3.1c}$$

where $\delta \ll \epsilon \ll 1/C = \mathcal{O}(1)$ and the single currents $I_\alpha(v, \mathbf{w}, c)$ are each of the type (2.6) and $\alpha \in \{\text{Na}, \text{K}, \text{Ca}, \text{Cl}, \text{L}, \text{KCa}\}$.[3] As noted by Rinzel and Ermentrout [37], there are several mechanisms which could provide the slow negative feedback required for bursting, in which c cycles periodically, causing transitions between fixed points and limit cycles in the fast (v, m) subsystem (see also [43]). For simplicity, we choose a very slow persistent potassium current (essentially the I_{KS} of Butera, Rinzel, and Smith [17]; see Model 2 in Appendix C and cf. [14, 17, 32]), but the results adapt to other mechanisms as shown in the example of section 4.3.

Given that the remaining gating variable m is slow relative to the voltage (spike) timescale, while c evolves yet more slowly, (3.1) has three time scales: the "fast" gating variables implicit in I_α, with $w_i = n_i(v)$ where appropriate, evolve on scales of order 1, the slower variable m

[3] In the following an additional subscript $()_f$ or $()_s$ will be added, e.g., K_f or K_s, to distinguish between fast and slow currents specific to a particular ion (here K).

evolves on a scale of order ϵ, and the very slow c dynamics has order δ. Such singularly perturbed systems have the appeal that the dynamic evolution can be separated according to the disparate time scales [36]. In the present case, currents can be divided into three groups: fast, instantaneously equilibrated currents, dynamic currents that evolve on the time scale of interest, and very slow variables that may be regarded as pseudostationary. The resulting simplification provides insight into the influence of single currents as part of a larger group.

3.1. The fast subsystem. We first analyze the fast subsystem (3.1a)–(3.1b). Since it varies very slowly, c will initially be treated as fixed, its dynamical effects being addressed subsequently. Hypothesis H1 implies that we can group the fast variables in (3.1a), write them as $I_{fv}(v)$, and separate them from the slow current of the form $I_s(v, m) = \sigma_{sm}(m)I_{sv}(v)$ (see, e.g., the Rose–Hindmarsh and Butera–Rinzel–Smith models [14, 17] given in Appendices A and C). This factorization is always possible for a single slow current by hypothesis H2 and extends to two (or more) provided that $\sigma_{\alpha_1}(w) = \sigma_{\alpha_2}(w)$; cf. [14] and Appendix A. Thus we can write (3.1a)–(3.1b) more explicitly as

$$C\dot{v} = -[I_{fv}(v) + \sigma_{sm}(m)I_{sv}(v)] + I_{\text{ext}}, \tag{3.2a}$$

$$\dot{m} = \frac{\epsilon}{\tau_m(v)}[m_\infty(v) - m]. \tag{3.2b}$$

In (3.2a) the subscripts sm and sv reflect functional dependence on the (slow) gating variable m and voltage v; note that the former enters only via $\sigma_{sm}(m)$. The voltage-dependent fast and slow currents I_{fv} and I_{sv} are given by

$$I_{fv}(v) = \sum_i \bar{g}_{\alpha_i}\sigma_i(n_i(v)) \cdot \zeta_i(v)(v - E_{\alpha_i}), \tag{3.3}$$

$$I_{sv}(v) = \bar{g}_{\alpha_s}\zeta_s(v)(v - E_{\alpha_s}). \tag{3.4}$$

As argued in section 2, the slow current gating variable enters (3.2a) as

$$\sigma_{\text{P}}(m) = m \quad \text{or} \quad \sigma_{\text{T}}(m) = m(1 - m). \tag{3.5}$$

We will call these cases *dynamically* persistent and *dynamically* transient, respectively, adding the term "dynamically" because persistent and transitory usually refer to steady state properties (cf. Figure 1); we are concerned here with currents whose dynamical dependence on the slow gating variable w makes them *appear* persistent or transitory.

The functions $\zeta_i(v)$ can often be assumed constant, but in some cases this does not suffice. A common counterexample is a transitory conductance with one gating variable significantly faster than the other, e.g., a fast activating and slowly inactivating sodium current $I_{\text{NaP-h}}$ [17] (Model 1 in Appendix C). In this case, setting the fast n_i at steady state, the conductance can be expressed in the form (3.4), with $\zeta(v) = n_{1_\infty}(v)$. Such currents should properly be called "mixed," since they are dynamically persistent, having the form $\sigma_{sm}I_{sv}$ in (3.2a), but appear transitory at steady state, due to the product of two gating variables.

3.1.1. Fixed points: One current. We now analyze the effect of the ionic currents in (3.2) on the location of fixed points of the fast subsystem. We begin with single currents.

We start by noting that separation into fast and slow currents (cf. (3.2)) has no influence on the location and number of fixed points because at such points all gating variables and currents are in equilibrium and can be expressed in terms of the voltage-dependent functions $n_{i\infty}(v)$ and $m_\infty(v)$.[4] The fixed points are therefore completely determined by the zeros of the right-hand side of (3.2a), called the $I_{ss} - v$ curve, sometimes written $I(v)$; this is the function measured in a voltage-clamp experiment. Moreover, we need only consider the case $\zeta(v) = 1$. Indeed, when this does not hold, the results are similar to the transitory current case, because the nonlinearity in $\zeta(v)$ acts as an additional linear or exponential multiplicative term, which does not change the qualitative form of $I(v)$.

Generic currents are described by

$$I(v) = \bar{g} \cdot \sigma(v; k_0, v_{th}) \cdot (v - E) \tag{3.6}$$

and depend upon four parameters: $\bar{g}$, k_0, v_{th}, and E. Figure 2 shows typical examples of the dependence on these. In [47] we show how all known currents fall into one of the four classes above. Here σ is of either form in (3.5), and the gating variable w is set to equilibrium, i.e., $w = w_\infty(v)$ (cf. (2.3)). We note the following.

The *maximal conductance* $\bar{g}$ acts as a scaling factor, affecting the values of critical points and their locations.

The *Nernst potential* E fixes the unique value of voltage $v = E$ for which the current vanishes. For transitory conductances, the current asymptotically approaches zero as $v \to \pm\infty$, but only as $v \to -\infty$ for persistent conductances. E also affects locations and values of the extrema.

The *threshold voltage* v_{th} affects locations and values of extrema. For transitory conductances they approximately coincide with the voltage that globally minimizes (maximizes) the current for $v_{th} < E$ ($v_{th} > E$). When $v_{th} > E$, the current is "essentially" monotonically increasing for physiological values of k_0. For persistent currents, the relative location of the threshold voltage v_{th} with respect to the reversal potential E can substantially influence the shape of $I(v)$. When $v_{th} < E$ (subreversal threshold), I has a distinct shape with a pronounced minimum (Figure 2a), e.g., I_{Na} in the HH equations [2]. If $v_{th} > E$ (superreversal threshold), the minimum is negligible, e.g., the potassium current I_{K} in the HH equations [2] (Figures 2c, d).

The *slope* k_0 determines the extent of the transition region from the inactive state $I \approx 0$ to the active state. For very small values of k_0, the currents tend to be linear over a wide range. In the limit $k_0 \to \infty$ the currents approach piecewise linear functions, and transitory currents are nonzero over only a very narrow range.

The substantial dips evident in Figures 2a, b are of particular importance in practice, since they imply regions of negative resistance characteristic (NRC) in the steady state $I_{ss} - v$ curves. As recognized experimentally by Wilson and Wachtel [48] in 1974, this is a necessary condition for bursting. We may anticipate that it is also necessary for Hopf (H) bifurcations. It is appreciable only when the threshold voltage is less than the Nernst potential, i.e., $v_{th} < E$. Since persistent currents play an important role, we conclude by noting that in the subreversal case and for slopes higher than a critical value $k_0 > k_{cr} = \frac{2}{E - v_{th}}$ that is usually exceeded in

[4] For clarity, we drop the subscript α but recall that each current comes with its own set of four parameters.

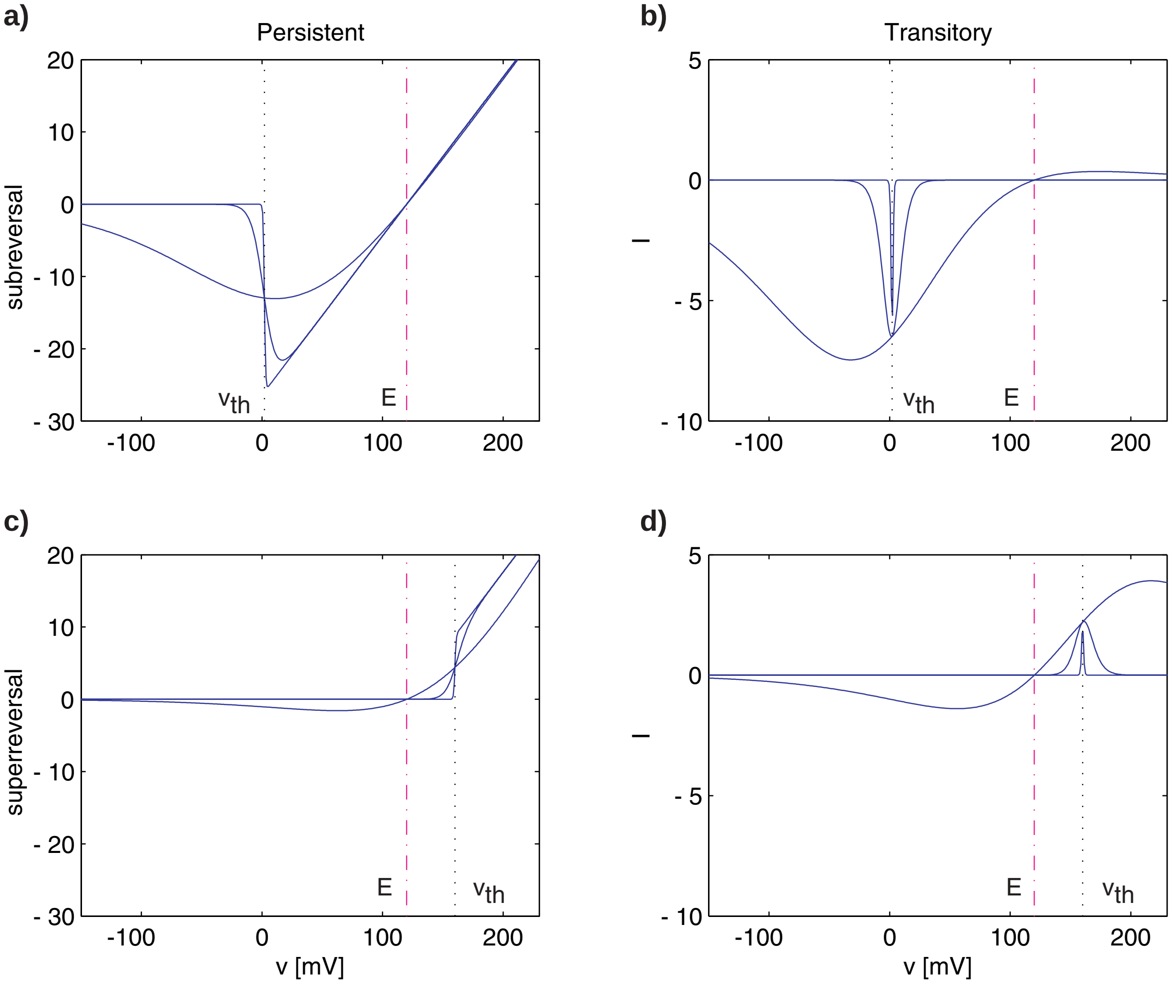


Figure 2. *Persistent (left) and transitory (right) ionic currents for different values of the slope parameter* $k_0 = 0.02$, 0.2, *and* 2 *(a typical value is* 0.1, *e.g.,* [2, 38]*). We illustrate with a calcium current of the form* (3.6)*, with maximal conductance* $\bar{g}_{Ca} = 4.4\,\mathrm{mS/cm}^2$ *and reversal potential* $E_{Ca} = 120\,\mathrm{mV}$ *and show the* $I_{ss} - v$ *curves* (3.6) *for the two conductance cases of* (3.5)*. Dotted and dash-dotted vertical lines show the threshold voltage* v_{th} *and reversal potential* E*.* (a) *Persistent subreversal* $v_{th} = 0\,\mathrm{mV} < E$*.* (b) *Transitory subreversal.* (c) *Persistent superreversal* $v_{th} = 160\,\mathrm{mV} > E$*.* (d) *Transitory superreversal. Note that transitory currents typically exhibit smaller ranges than persistent currents (≈ 20%; note differing scales on ordinate* I*).*

physiological ranges, the minimum is bounded below by $I_{\min} \geq \bar{g} \cdot (v_{th} - E)$. Finally, we note that *passive conductances* or passive currents, like the leakage $I_{\mathrm{L}} = g_{\mathrm{L}} \cdot (v - E_{\mathrm{L}})$ can be described as a degenerate subclass of persistent currents with zero slope: $k_0 = 0$.

The current I_{ext} enters (3.2a) as a purely additive term, so for any voltage $v = \bar{v}$, one can find a current I such that the fixed point is at $\bar{v}$. Hence only the general shape of the $I_{ss} - v$ curve is relevant in determining the possible number of fixed points. We may therefore conclude that *existence of an NRC “dip” can introduce up to two new fixed points.*

3.1.2. Fixed points: Multiple currents.

Linear or passive currents. In the absence of leakage or other linear currents, the existence of at least one fixed point, typically at low voltage values, is no longer guaranteed. Apart from this, passive currents (with positive conductance) cannot generate, but only destroy, fixed points (see discussion below).

Nonlinear currents. As described above, the most relevant feature is the creation of local minima in the $I_{ss} - v$ curve, but if $v_{th} > E$, the resulting dips are negligibly small (Figures 2c, d). We will therefore consider the subreversal case $v_{th} < E$. For simplicity we discuss only persistent currents, but one can also give bounds for transitory ones. For equilibrated gating variables $w = n_\infty(v)$ of the form (2.3) the persistent currents and their (voltage) derivatives are

$$\begin{aligned} I_{\mathrm{P}} &= \bar{g} \cdot n_\infty \cdot (v - E), \\ I'_{\mathrm{P}} &= \bar{g} \cdot n_\infty \cdot [k_0(1 - n_\infty)(v - E) + 1], \\ I''_{\mathrm{P}} &= \bar{g}k_0 \cdot n_\infty \cdot (1 - n_\infty)[k_0(1 - 2n_\infty)(v - E) + 2]; \end{aligned} \tag{3.7}$$

hence the minimum occurs at

$$I_{P_{\min}} = \bar{g}\left(\bar{v} - E + \frac{1}{k_0}\right), \tag{3.8}$$

where $\bar{v}$ is implicitly defined by $(1 - n_\infty(\bar{v}))(\bar{v} - E) = -\frac{1}{k_0}$. The addition of a current can destroy the local "dip" of a pre-existing current. A sufficient condition for this is that the derivative of the new current be larger in magnitude than the pre-existing one; if the added current always increases more than the other decreases, no local minimum survives. It is therefore useful to estimate the maximum slope of I_P, which is obtained at its inflexion point to the left of E:

$$I'_{P_{\min}} = \frac{\bar{g}}{4}[k_0(\tilde{v} - E) + 2]. \tag{3.9}$$

Here the voltage $\tilde{v}$ is implicitly defined by $(1 - 2n_\infty(\tilde{v}))(\tilde{v} - E) = -\frac{2}{k_0}$. As anticipated, $I_{P_{\min}}$ is bounded below by $\bar{g}(v_{th} - E)$, achieved in the limit $k_0 \to +\infty$. In the same limit, the minimum derivative is unbounded and tends to $-\infty$. Therefore, any nonlinear current can create up to two new fixed points.

One can show this in general; in particular, consider the limit of high thresholds $k_{0_i} \to \infty$ for all i, and let the individual voltages be ordered as $v_{th_1} < E_1 < v_{th_2} < E_2 < \cdots < v_{th_N} < E_N$. In this limit one can define "influence windows" $U_i = [v_{th_i}, E_i]$ such that $n_j(v) \approx 0$ or 1 for all $j \neq i$; i.e., in the ith window only the current I_{α_i} is "turning on or off"; the others are all inactive or fully active. Suppose further that $I_j = I_{P_j} = \bar{g}_{\alpha_j} n_j(v)\,(v - E_{\alpha_j})$ is persistent. Then, it follows that the total current and its derivative are

$$I = \sum_{i \neq j} \bar{g}_{\alpha_i}(v - E_{\alpha_i}) + I_{P_j} \quad \text{and} \quad I' = \sum_{i \neq j} \bar{g}_{\alpha_i} + I'_{P_j}. \tag{3.10}$$

(In (3.10) we set $\bar{g}_{\alpha_i} = 0$ for inactive currents.) From (3.9), I'_{P_j} can be arbitrarily large and negative, and analogous arguments hold for transitory currents. We note that, for increasing

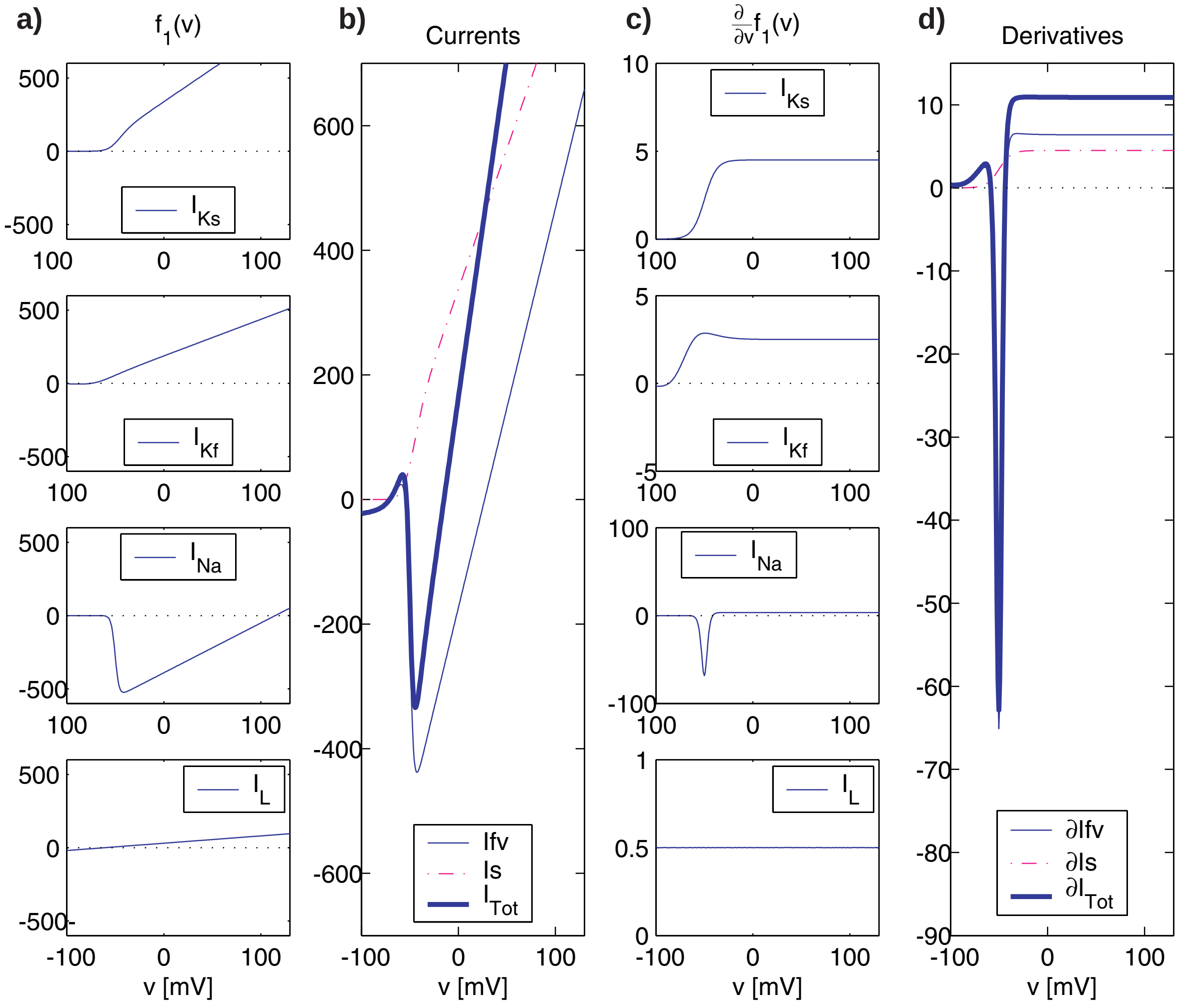


Figure 3. *An example of four ionic currents: a slow potassium current I_{Ks}, two fast potassium and sodium currents I_{Kf} and I_{Na}, and a leakage current I_L. Parameters are as follows: $\bar{g}_{Ks} = 4.5$, $E_K = -75$, $v_{th_{Ks}} = -50$, $k_{0_{Ks}} = 0.08$; $\bar{g}_{Kf} = 2.5$, $v_{th_{Kf}} = -70$, $k_{0_{Kf}} = 0.05$; $\bar{g}_{Na} = 3.4$, $E_{Na} = 115$, $v_{th_{Na}} = -50$, $k_{0_{Na}} = 0.25$; $\bar{g}_L = 0.5$, $E_L = -60$, with units as given in section 4. Leftmost panels* (a) *show the $I_{ss} - v$ curves for the individual ionic currents.* (b) *The fast currents are collected in I_{fv} and are shown solid, the slow current I_{sv} is shown dash-dotted, and the sum $I_{Tot} = I_{ion}$ is shown bold. In* (c) *the derivatives of the single ionic currents are shown, and* (d) *shows the derivative with respect to v of the collected fast currents (dashed), the derivative of the slow current (dash-dotted), and their sum (bold), which gives the coefficient $a = \frac{I_{ion}}{\partial v}$ of section 3.1.3.*

numbers of currents N, $\sum_i^N \bar{g}_{\alpha_i}$ tends to increase since constant terms g_{α_i} are added which activate in the sequence $v_{th_1}, \ldots, v_{th_N}$. Therefore, the more currents there are, the less likely it is that they produce new fixed points, unless their conductances or slopes are very large.

Figure 3 shows an example with four ionic currents: two fast, I_{Kf} and I_{Na}; one slow, I_{Ks}; and a leakage current I_{L}. The leftmost panels show the steady state $I_{ss} - v$ curves for the individual currents, which are then collected in I_{fv} and I_s and added to give the total current I_{ion} shown in Figure 3b.

Conclusion. From the above analysis, we can summarize the major result as follows. *Each additional nonlinear current can, for suitable $\bar{g}, k_0, v_{th}, E$, create a new local "dip" and hence two additional fixed points may arise; cf.* [31]. *Leakage currents guarantee one fixed point for any value of the applied current.*

Here we have emphasized the role of the $I_{ss} - v$ curve in determining fixed points, due to its biophysical relevance. Indeed, this is the characteristic measured in voltage-clamp experiments, and can therefore be directly related to data. Alternatively, fixed points may be found at intersections of the v- and m-nullclines of (3.2) (e.g., [49, 22, 37, 33]).

3.1.3. Stability.

In this section, we analyze the stability of fixed points. We concentrate on slow dynamically persistent currents, which are more common in reduced models, but we also discuss slow transitory currents. Rewriting (3.2) as

$$\begin{aligned} \dot{v} &= -\frac{1}{C}[I_{fv}(v) + \sigma_{sm}(m)I_{sv}(v) + I_{\text{ext}}] \overset{\text{def}}{=} f_1(v,m), \\ \dot{m} &= \frac{\epsilon}{\tau_m(v)}[m_\infty(v) - m] \overset{\text{def}}{=} \epsilon f_2(v,m) \end{aligned} \tag{3.11}$$

and linearizing yield a Jacobian of the form

$$Df = \begin{bmatrix} -\frac{1}{C}\,a & -\frac{1}{C}\,b \\ d\epsilon & -e\epsilon \end{bmatrix}. \tag{3.12}$$

If we define the total ionic current $I_{\text{ion}} = -(I_{fv} + \sigma_{sm} I_{sv})$, then the coefficients evaluated at a fixed point p are given by

$$a = \frac{\partial I_{\text{ion}}}{\partial v}\,|_p, \quad b = \frac{\partial I_{\text{ion}}}{\partial m}\,|_p, \quad d = \frac{m'_\infty}{\tau_m} - \frac{(m_\infty - m)\tau'_m}{\tau_m^2}\,|_p, \quad e = \frac{1}{\tau_m}\,|_p\,. \tag{3.13}$$

Here a represents the variation of the ionic current with respect to voltage, sometimes called the instantaneous $I - v$ curve [37] or the *slope conductance curve* [41]. The coefficient b reflects the dependence of the ionic current on the slow variable, and d and e are entirely determined by the gating dynamics. Observing that the sigmoid (2.3) has the property that its derivatives can be expressed in terms of the function itself, e.g., $w'_\infty = k_0 w_\infty(1 - w_\infty)$, we may write

$$\begin{aligned} a &= I'_{fv} + \sigma_{sm} I'_{sv}, \quad b = \frac{\partial \sigma_{sm}}{\partial m} I_{sv}, \\ d &= k_0 \frac{m_\infty(1 - m_\infty)}{\tau_m}, \quad e = \frac{1}{\tau_m}, \end{aligned} \tag{3.14}$$

where $(\cdot)' = \frac{\partial}{\partial v}(\cdot)$ and the derivatives of the conductance factors σ_{sm} are given by $\frac{\partial \sigma_P}{\partial m} = 1$ and $\frac{\partial \sigma_T}{\partial m} = 1 - 2m$ for dynamically persistent and dynamically transitory conductances, respectively (cf. (3.5)). In computing d we note that the second term in the general expression of (3.13) vanishes at fixed points. Also note that d and e are always positive.

As noted above, the particular structure of (2.2) implies that at the fixed points all gating variables are explicit functions of voltage. In addition, and importantly, as we noted at the end of section 3.1.1, any voltage value $v = \bar{v}$ can be made a fixed point by suitable choice of external current I_{ext}. Therefore, the Jacobian entries a, b, d, e of (3.13) can all be reduced to explicit functions of voltage at the fixed point $\bar{v}$. This substantially simplifies the stability analysis, reducing it to a characterization in terms of $\bar{v}$ alone.

The eigenvalues of (3.12) are determined by the determinant DetDf and trace TrDf and the necessary conditions for H and saddle-node (SN) bifurcations [50] may be written

$$a_H = -\epsilon e C, \quad ae + bd > 0, \tag{3.15a}$$

$$a_{SN} = -\frac{bd}{e}. \tag{3.15b}$$

Using (3.14), we observe that the ratio $-\frac{d}{e}$ appearing in (3.15b) is given by

$$-\frac{d}{e} = -k_0 m_\infty (1 - m_\infty) \tag{3.16}$$

and depends only on the (slow) gating dynamics; it is affected neither by the addition of fast currents nor by whether the slow current is dynamically persistent or transitory, activating or inactivating. In addition it depends neither on the maximal conductance $\bar{g}_\alpha$ nor on the reversal potential E_α, but only on the slope k_{0_α} and threshold voltage v_{th_α}. It is a negative bell-shaped function tending exponentially to 0^- for $v \to \pm\infty$; e.g., see Figure 5c. We can therefore focus on the coefficients a and b.

Coefficient a. The slope conductance curve is composed of the terms I'_{fv} and $\sigma_{sm} I'_{sv}$ (cf. (3.14)). It is often stressed [41, 39] that stability cannot be inferred from the slope of the $I_{ss} - v$ curve. Indeed, $I'_{ss} = \frac{\partial I_{\text{ion}}}{\partial v}$ would be equal to a if all currents were fast, but in the presence of slow currents, this is no longer true. Gathering N fast currents, using (3.3), we have

$$\frac{\partial I_{\text{ion}}}{\partial v} = I'_{fv} = \sum_{i=1}^{N} \bar{g}_{\alpha_i} \sigma'_i \zeta_i(v)(v - E_{\alpha_i}) + \bar{g}_{\alpha_i} \sigma_i \zeta'_i(v)(v - E_{\alpha_i}) + \bar{g}_{\alpha_i} \sigma_i \zeta_i(v), \tag{3.17}$$

where $\sigma'_i = \frac{\partial \sigma(n_i(v))}{\partial v}$. However, if one of the currents is slow, then we have

$$\begin{aligned} \frac{\partial I_{\text{ion}}}{\partial v} = & \sum_{i=1}^{N-1} \left[\bar{g}_{\alpha_i} \sigma'_i \zeta_i(v)(v - E_{\alpha_i}) + \bar{g}_{\alpha_i} \sigma_i \zeta'_i(v)(v - E_{\alpha_i}) + \bar{g}_{\alpha_i} \sigma_i \zeta_i(v) \right] \\ & + \bar{g}_{\alpha_j} \sigma_j \zeta'_j(v)(v - E_{\alpha_j}) + \bar{g}_{\alpha_j} \sigma_j \zeta_j(v), \end{aligned} \tag{3.18}$$

and the analogue of the first term $\bar{g}_{\alpha_i} \sigma'_i (v - E_{\alpha_i})$ in the summation does not appear for the slow current j.

Figure 4 shows the case of $\zeta_i(v) = 1$, in which (3.18) simpifies to

$$\frac{\partial I_{\text{ion}}}{\partial v} = \sum_{i=1}^{N-1} \left[\bar{g}_{\alpha_i} \sigma'_i (v - E_{\alpha_i}) + \bar{g}_{\alpha_i} \sigma_i \right] + \bar{g}_{\alpha_j} \sigma_j. \tag{3.19}$$

Since $\sigma'_{\text{P}} = k_0 n(1 - n)$ and $\sigma'_{\text{T}} = k_0 n(1 - n)(1 - 2n)$ (from (2.3) and (3.5); cf. (3.7)), the first term in the sum is a hump or a "dipole" for dynamically persistent or transitory currents,

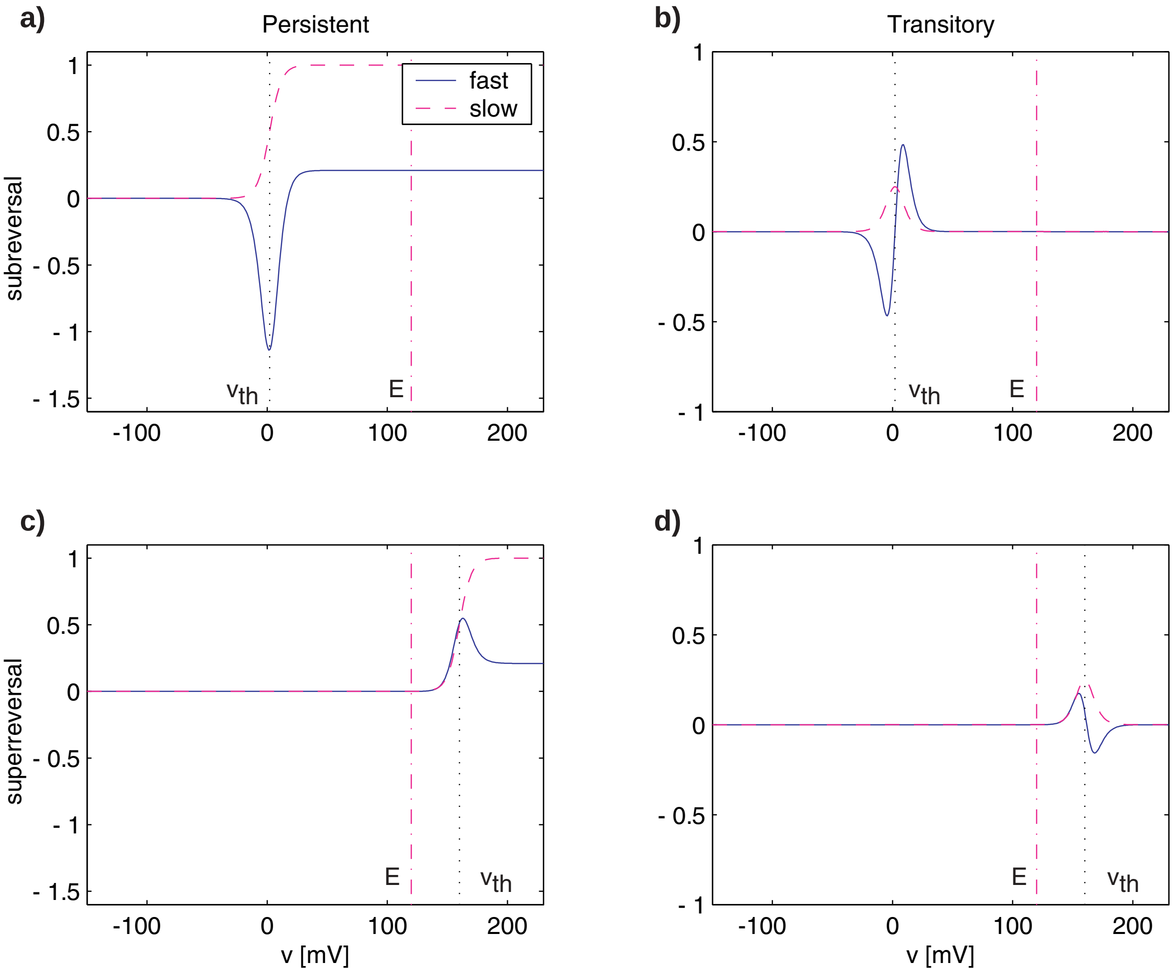


Figure 4. *Derivatives of persistent ionic currents (left) and transitory currents (right), showing how they differ for fast (solid) and slow (dashed) currents. See Figure 2 for steady state $I_{ss} - v$ curves. This example shows a calcium current of the form* (3.6)*, with maximal conductance $\bar{g}_{Ca} = 4.4\,\mathrm{mS/cm^2}$, and reversal potential $E_{Ca} = 120\,\mathrm{mV}$, with $k_0 = 0.2$. Dotted vertical line shows the threshold voltage v_{th}; dash-dotted vertical line shows the reversal potential E.* (a) *Persistent subreversal $v_{th} < E$, here $v_{th} = 0\,mV$.* (b) *Transitory subreversal.* (c) *Persistent superreversal $v_{th} > E$, here $v_{th} = 160\,mV$.* (d) *Transitory superreversal. Observe that slow currents always give positive contributions to the derivative (coefficient a).*

respectively. If current j is slow, then the derivative $\frac{\partial I_\alpha}{\partial v} > 0$ for both types of currents, but if j is fast, then $\frac{\partial I_\alpha}{\partial v}$ can change sign. In the persistent case, this will only happen for subreversal currents (Figures 4a and 4c), whereas it always holds in the transitory case (Figures 4b and 4d).

Reviewing the four-current example of Figure 3, we observe how the individual currents contribute to determine $a = \frac{\partial I_{\text{ion}}}{\partial v}$ depicted in Figure 3d. I_{Kf} is a typical example of a persistent superreversal current whose derivative is shown in Figure 3c (second down). I_{Na} exemplifies a fast subreversal persistent current whose derivative is shown in Figure 3c (third down); cf. I_{NaP} in [17], Model 2 of Appendix C. The linear leak I_{L} gives a constant contribution to a; see

Figure 3c (bottom). I_{Ks} is a slow (dynamically and statically) persistent current; see Figure 3c (top). Note the difference between the two potassium currents. The slow current I_{Ks}, also present in the HH equations [2], yields a term $a_{\mathrm{Ks}} = g_{\bar{\mathrm{K}}\mathrm{s}} m_{\infty_{\mathrm{Ks}}}$ (Figure 3c$_1$), whereas the fast current I_{Kf} yields the term $a_{\mathrm{Kf}} = \bar{g}_{\mathrm{Kf}}\, k_{\mathrm{Kf}}\, n_{\infty_{\mathrm{Kf}}}\,(1 - n_{\infty_{\mathrm{Kf}}})(v - E_{\mathrm{K}}) + \bar{g}_{\mathrm{Kf}} n_{\infty_{\mathrm{Kf}}}$ (Figure 3c (second down)). The difference is not very marked here, but, as shown in Figure 4a, it can be more substantial and sufficient to change the stability of a fixed point. The single terms $\frac{\partial I_\alpha}{\partial v}$ are constant for $v \to \pm\infty$.

Coefficient b. This coefficient describes the dependence of the ionic current upon the (relatively) slow gating variable m. For dynamically persistent and transitory currents, one obtains

$$b_{\mathrm{P}} = I_{sv} \quad \text{and} \quad b_{\mathrm{T}} = (1 - 2m_\infty) I_{sv}. \tag{3.20}$$

For dynamically persistent currents b_{P} has the same sign as I_{sv}; hence if, as usually, $I_{sv} = v - E_{\alpha_s}$ is linear, then b_{P} is strictly positive for all $v > E_{\alpha_s}$.

It is now relatively easy to analyze the behavior of fixed points. One computes the dependence of a on $\bar{v}$ at a fixed point; when a crosses one of the values a_{SN} or a_H defined by (3.15), stability changes and a bifurcation occurs. An example is shown in Figure 5 for a system with the currents of Figure 3. The condition $\mathrm{Tr} = 0$, satisfaction of which with $\mathrm{Det} > 0$ results in an H bifurcation, is depicted in Figure 5a. Note that the term $-e\epsilon$ is small only in a relatively narrow range of voltages because $e(v) = \frac{1}{\tau_m(v)} \to +\infty$ for $v \to \pm\infty$; asymptotic analysis is therefore of little help for global understanding. Figure 5c shows the term $-\frac{d}{e}$, which multiplied by b_{P} or b_{T} gives the condition $\mathrm{Det} = 0$. For persistent currents, the SN condition a_{SN} is shown in Figure 5b. Since a_{SN} is always positive for $v < E_{\mathrm{Ks}}$, the SN and the H bifurcations can never occur in that range, but only at more depolarized levels than the reversal potential of the slow variable, here E_{Ks}. We note that since ϵ will only change the shape of $\mathrm{Tr} = 0$ (by flattening it), the above observation suggests that there is a lower bound for these bifurcations and, as $\epsilon \to 0$, the location of these points will not change much. The determinant is positive for $a > -b\frac{d}{e}$, above the bold lines in Figures 5b and 5d for dynamically persistent and transitory conductances, respectively. Finally, Figure 5e shows the two boundaries with a superimposed for the example of Figure 3. This reveals a first crossing of a_H for $\bar{v}_{H_1} \approx -58\,m\mathrm{V}$ giving rise to an H bifurcation, followed by two intersections of a_{SN} at $\bar{v}_{SN_1} \approx -57.8\,m\mathrm{V}$ and $\bar{v}_{SN_2} \approx -44.3\,m\mathrm{V}$ and finally a second H bifurcation at $\bar{v}_{H_2} = -43.8\,m\mathrm{V}$, giving the bifurcation sequence H, SN, SN, H. In the next section we will exploit this approach to explore the effect of parameter variations.

3.2. Bifurcation diagrams for the fast subsystem. The collective effect of parameters describing single currents are best exemplified in bifurcation diagrams. The discussion of section 3.1.3 immediately translates to a bifurcation diagram with external current I_{ext} as the bifurcation parameter. The diagrams given below were computed numerically using a Newton–Raphson algorithm to determine fixed points and a continuation algorithm to follow their branches. However, the "constructive" single current analysis developed above more clearly reveals the causes and parameter sensitivities responsible for changes in the structure and sequence of bifurcations along branches of equilibria, so we also display this information in the form of $I_{ss} - v$ and slope conductance curves. In the following we will assume $\zeta(v) = 1$ for simplicity.

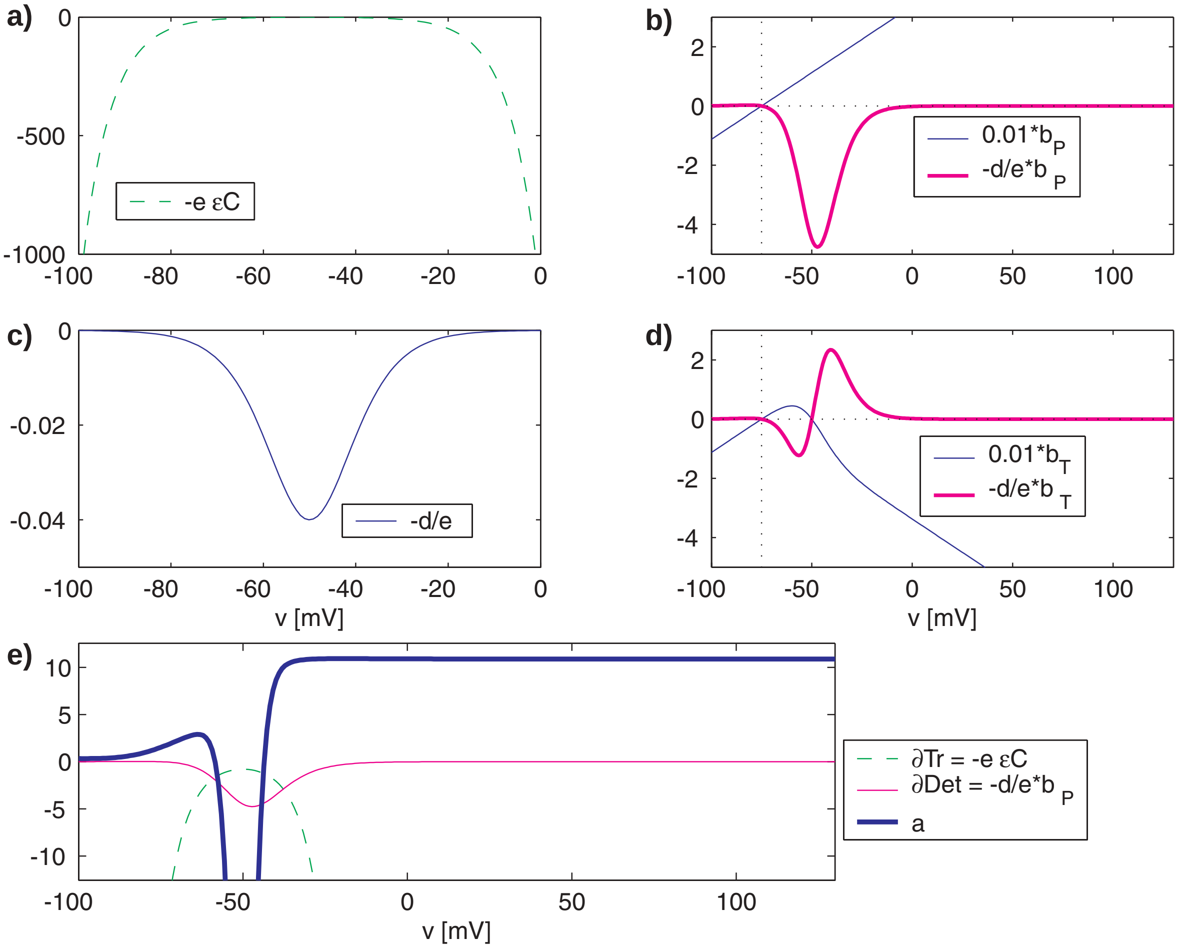


Figure 5. (a) *The H bifurcation condition $a_H = -\epsilon eC$* (3.15a)*: TrDf < 0 (> 0) above (below) dashed curve;* (b) *rescaled b_P (solid) and the SN bifurcation condition $a_{SN} = -b_P d/e$* (3.15b) *(bold);* (c) *the term $-d/e$;* (d) *rescaled b_T (solid) and the SN bifurcation condition $a_{SN} = -b_T d/e$ (bold).* (e) *Shows the H condition as in* (a) *(dashed), the SN condition as in* (b) *(solid), and the coefficient a (bold) for the system of four currents given in Figure* 3*. Other parameters are $C = 20$, $\epsilon = 0.04$. Note that voltage (v) scales differ.*

3.2.1. Fast currents. We consider a simple case with three ionic currents: a slow persistent potassium current I_{Ks}, a fast persistent calcium current I_{Ca}, and a leakage current I_{L}, similar to the original work of Hodgkin and Huxley [2].[5] The term

$$b\frac{d}{e} = k_0 m_\infty(1 - m_\infty) \cdot I_{sv} \cdot \begin{cases} 1, & \text{dynamically persistent,} \\ (1 - 2m_\infty), & \text{dynamically transitory,} \end{cases}$$

which gives the SN condition (3.15b), depends only on the slow gating parameters. The bifurcation sets (3.15) are therefore affected neither by adding fast currents nor by changes

[5] Sodium and calcium currents differ in their reversal potentials $E_{\mathrm{Na}} = 50m\mathrm{V}$ and $E_{\mathrm{Ca}} = 120m\mathrm{V}$ and in the fact that the sodium current in [2] is transitory, whereas the calcium current considered here is persistent, as in [38].

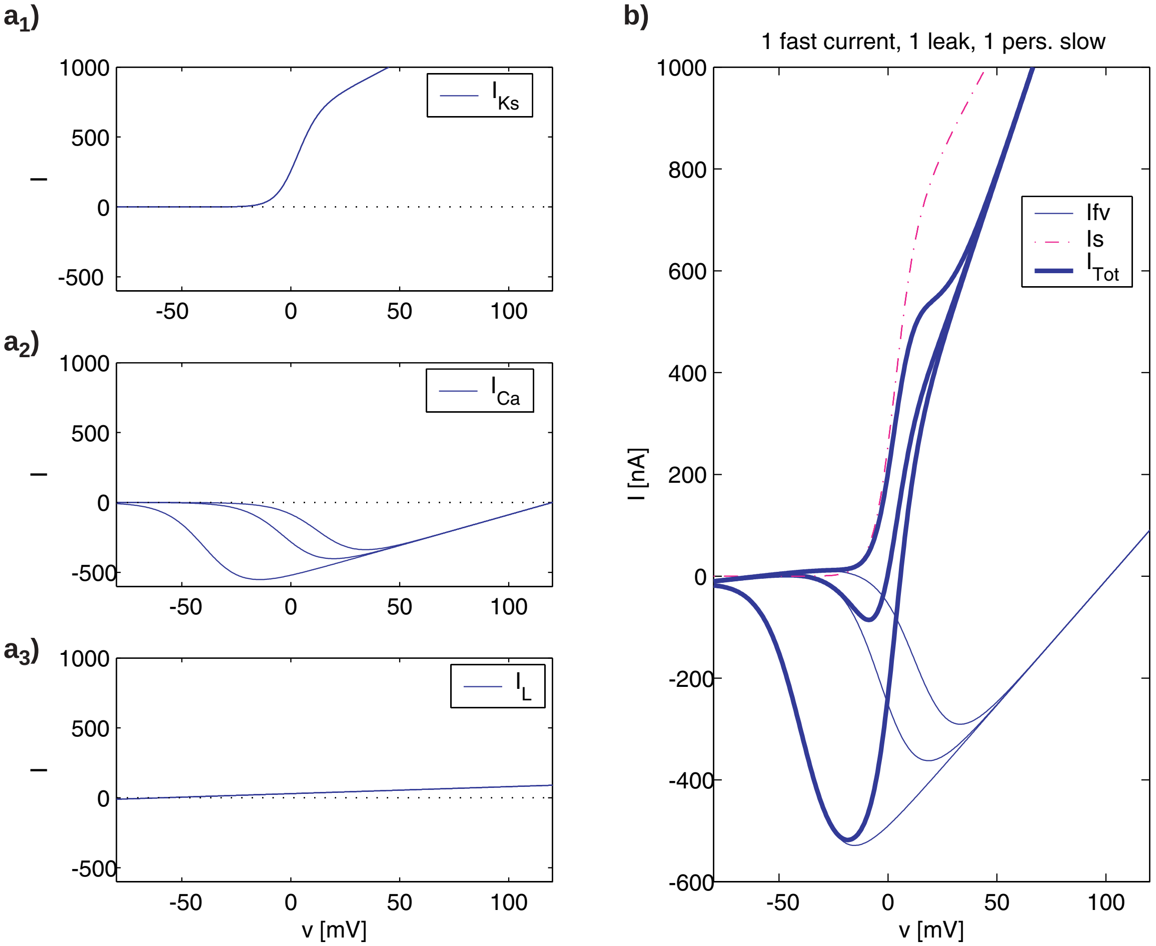


Figure 6. *Dependence of the total ionic current on variation of threshold voltage v_{th} of a fast calcium current. Parameters are as follows: $\bar{g}_{Ks} = 8.0$, $E_K = -80$, $v_{th_{Ks}} = 2$, $k_{0_{Ks}} = 0.2$; $\bar{g}_{Ca} = 4.4$, $E_{Ca} = 120$, $v_{th_{Ca}} = -38, -1.2, +15$ from lower to upper curves in* (a$_2$) *and* (b), *$k_{0_{Ca}} = 0.11$; $\bar{g}_L = 0.5$, $E_L = -60$. Units are as given in section* 4.

in their parameters. For illustrative purposes, we show the effect of two such parameters: the threshold voltage v_{th} of a fast persistent inward current such as $I_{\rm Na}$ or $I_{\rm Ca}$, and the slope of its fast gating variable k_0.

Threshold voltage v_{th}. The effect of v_{th} on the $I_{ss} - v$ curve is shown in Figure 6, its effect on the slope conductance curve (coefficient a) in Figure 7, and the resulting bifurcation diagrams in Figure 8. Increasing values of v_{th} shift the minimum of $\frac{\partial I_{\rm Ca}}{\partial v}$ to the right. For low thresholds, the corresponding bifurcation diagram has two SN points (see Figure 8c$_1$). Increasing v_{th}, an H bifurcation emerges from the higher (more depolarized) SN bifurcation point in a Takens–Bogdanov (TB) bifurcation [50] (see Figure 8c$_2$). Further increase causes the SN points to coalesce and disappear in a codimension two "cusp" bifurcation [50], leaving two H bifurcations (see Figure 8c$_3$ (cf. Rinzel and Ermentrout [37] and Koch [41] for discussions of the latter)).

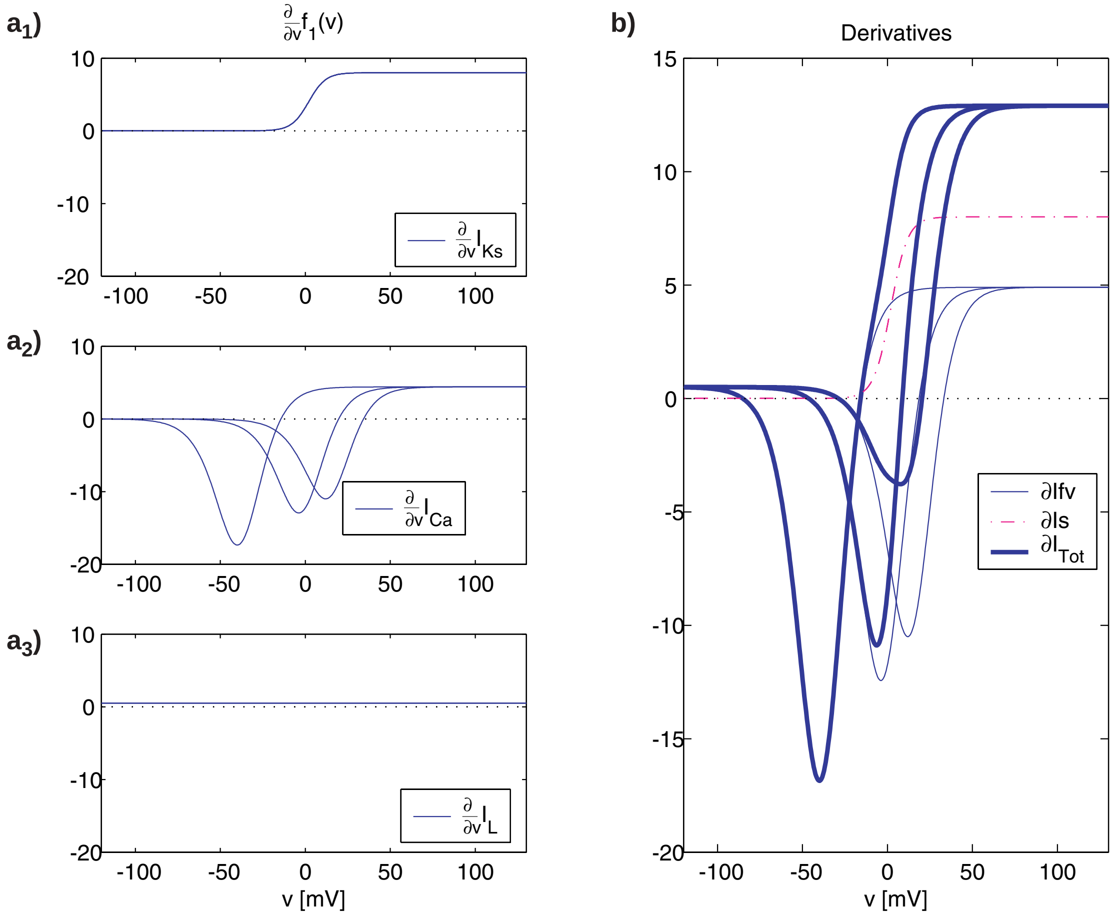


Figure 7. *Dependence of the total slope conductance curve a on variation of threshold voltage v_{th} of a fast calcium current. Parameters as in Figure 6.*

Also note that the cases of Figures 8a$_2$–c$_2$ and a$_3$–c$_3$ can explain the smooth transition from Class I to Class II spiking [20] without appealing to an extra current (e.g., an I_A current) as in Connor and Stevens [51], [41, pp. 159, 190]. Rinzel and Ermentrout [20] stated that this was possible with a model similar to the one used here by changing v_{th_K}; Figure 8 should provide some further insight.

Slope k_0. Despite the fact that a steeper transition in the sigmoid (2.3) has a negligible effect on the steady state curves (Figures 9a$_1$–a$_2$), it can substantially change the bifurcation structure via the increased slope that causes a substantial negative peak in $\frac{\partial I_{\mathrm{Ca}}}{\partial v}$ (see Figures 9b$_1$–b$_2$). Moreover, due to global bifurcations in which limit cycles disappear (see [50] and below), the topological difference between the two cases involves more than simply removing one (local) H bifurcation point (see Figures 9c$_1$–c$_2$).

Maximal conductance of leakage current $\bar{g}_L$. Because of its relevance to the bursting dynamics in the following section, we end by noting that the effect of the (linear) leakage current is simply a vertical shift of a. The resulting bifurcation diagram (not shown) goes from the sequence SN, SN, H to SN, SN as $\bar{g}_L$ increases.

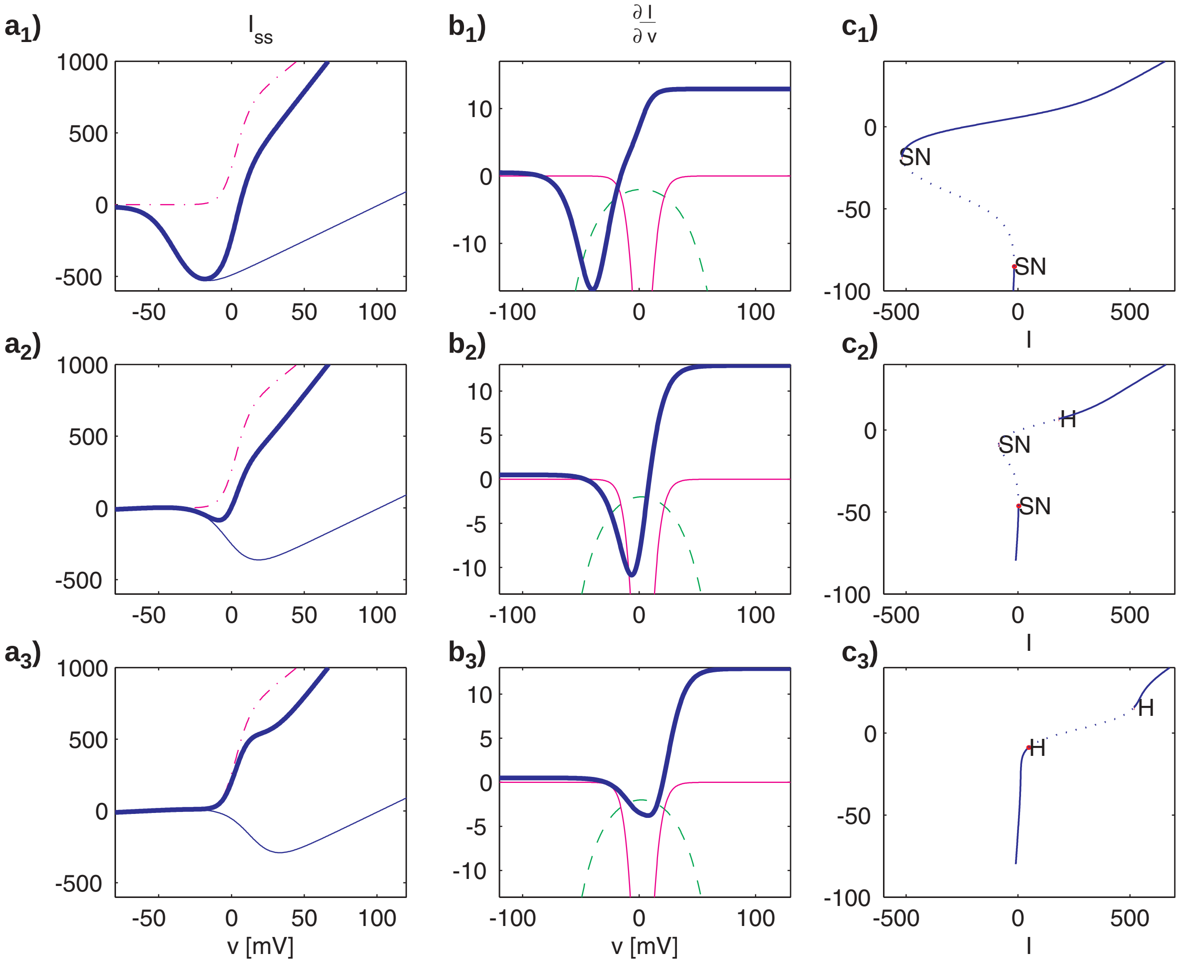


Figure 8. *Dependence of a fast calcium current upon variation of the threshold voltage $v_{th_{Ca}}$. Parameters as in Figure 6 with $v_{th_{Ca}} = -38, -1.2, +15$ from top row to bottom; in addition $C = 20$, $\epsilon = 0.1$. Left panels show the steady state $I_{ss} - v$ curves from Figure 6: fast current (solid), slow current (dash-dotted), and total current (bold). Middle panels show the relevant terms for stability from Figure 7: $Det = 0 \Leftrightarrow a = -\frac{bd}{e}$ (solid), $Tr = 0 \Leftrightarrow a = -e\epsilon C$ (dashed), and $a = \frac{\partial I_{ion}}{\partial v}$ (bold). Right panels show the corresponding bifurcation diagrams.*

3.2.2. Slow currents.

Slow current parameters also affect the bifurcation sets (3.15). Figure 10 shows the effect of threshold voltage changes on a slow outward current, such as I_{K}. Bifurcation points are shifted and the coefficient a changes its form via $\sigma_{sm}(m_{\infty}(v))$, which appears in the second term in $a = I'_{fv} + \sigma_{sm} I'_{sv}$. The resulting bifurcation diagrams show transitions can occur from SN, SN to SN, SN, H and back to SN, SN as v_{th} increases.

Conclusion. The introduction of each current with a nonoverlapping "window of influence" can produce another pair of equilibria. Thus "snaking" branches with multiple SN bifurcations can appear. Up to two H bifurcations can be introduced, associated with at least one SN pair. H bifurcations may also occur in the absence of SN bifurcations when the branch does not double back. Coincident H and SN (TB) bifurcations can be obtained by varying a second parameter in addition to I_{ext}.

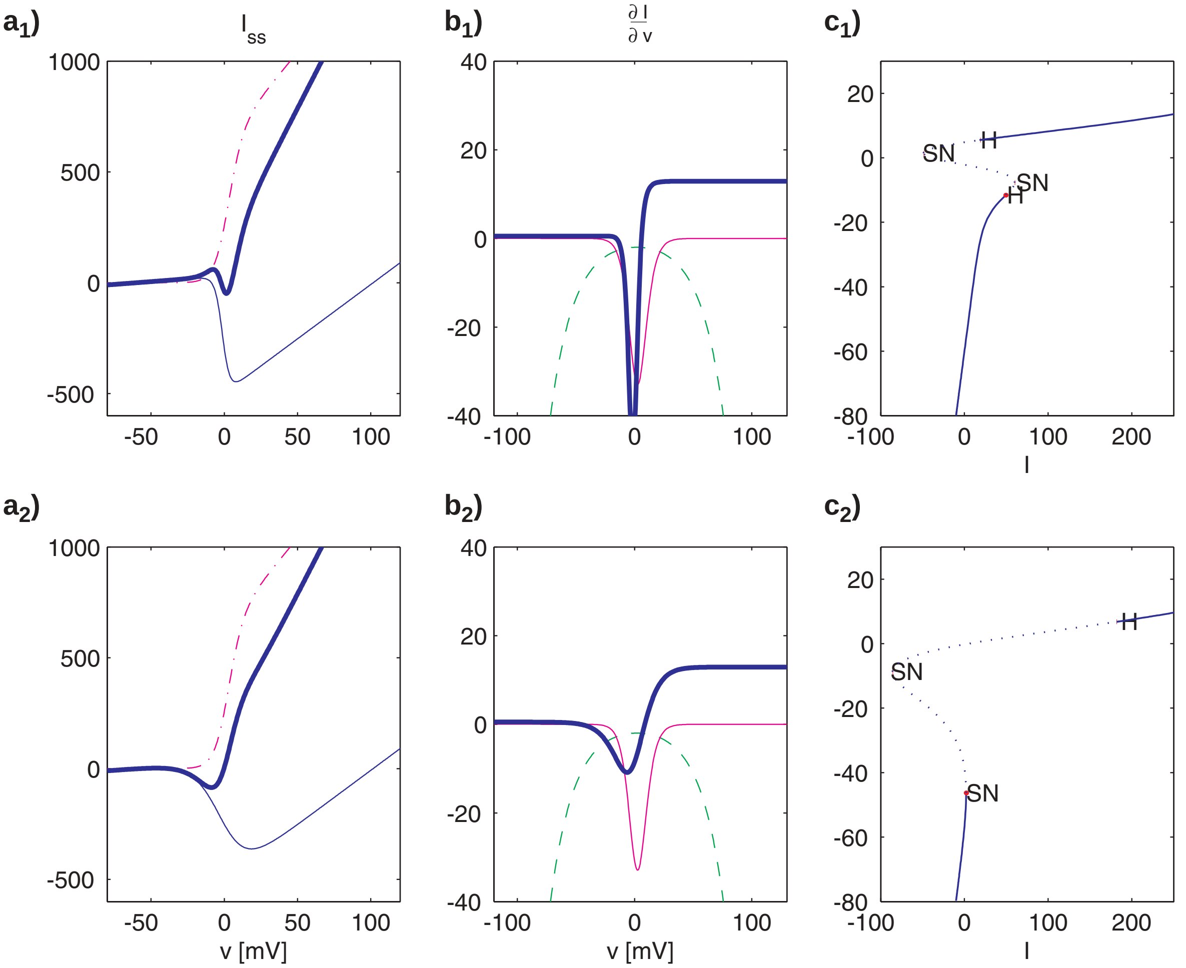


Figure 9. *Dependence of a fast calcium current upon variation of slope k_0. Left panels show the steady state $I_{ss} - v$ curves: fast current (solid), slow current (dash-dotted), and total current (bold). Middle panels show the relevant terms for stability: $Det = 0 \Leftrightarrow a = -\frac{bd}{e}$ (solid), $Tr = 0 \Leftrightarrow a = -\epsilon\epsilon C$ (dashed line), and $a = \frac{\partial I_{ion}}{\partial v}$ (bold). Right panels show the corresponding bifurcation diagrams.*

3.2.3. Bifurcations in terms of c. The bifurcation diagrams of Figures 8, 9, and 10 use external current I_{ext} as parameter. In the full system (3.1) the slow variable c drives the fast subsystem from regime to regime; hence, we must recast the above results in terms of c, which enters the fast equation (3.1a) via a current such as $I_{\text{KS}} = \bar{g}_{\text{KS}}c(v - E_{\text{K}})$. To do this we consider a two-parameter bifurcation diagram of the original system (3.1) and then slice it with an appropriate plane. For illustrative purposes, we will treat (3.1) with three internal currents $I_{\text{Ca}}, I_{\text{Ks}}, I_{\text{L}}$ and an external current I_{ext}, as in [38].

We compare the membrane voltage equations of (3.1a),

$$C\dot{v} = -\left[\bar{g}_{\text{L}}c_{\text{L}}(v - E_{\text{L}}) + f_1(v, m)\right] + I_{\text{ext}}, \tag{3.21}$$

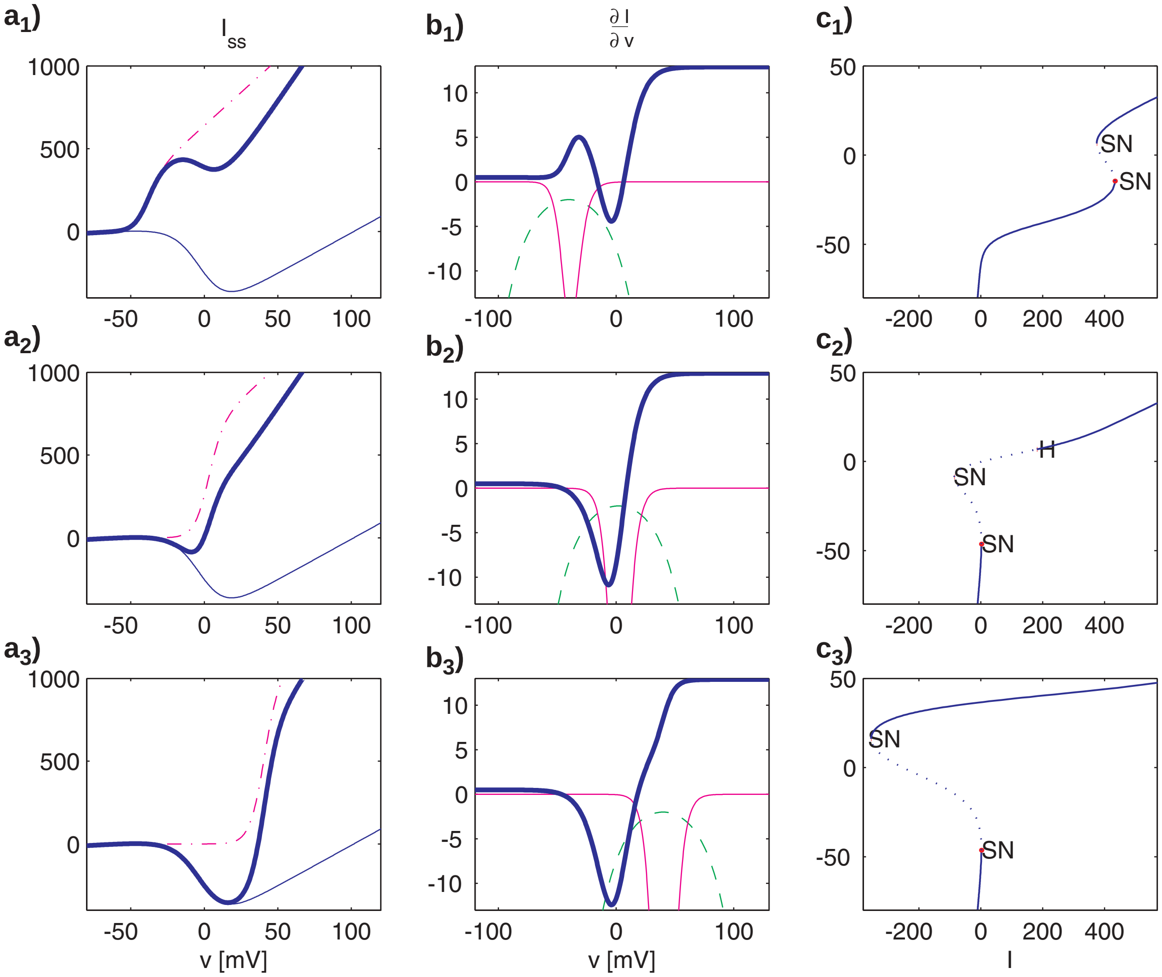


Figure 10. *Dependence of a slow current upon variation of the threshold potential v_{th} for a potassium current. Left panels show the steady state $I_{ss} - v$ curves: fast current (solid), slow current (dash-dotted), and total current (bold). Middle panels show the relevant terms for stability: $Det = 0 \Leftrightarrow a = -\frac{bd}{e}$ (solid), $Tr = 0 \Leftrightarrow a = -e\epsilon C$ (dashed), and $a = \frac{\partial I_{ion}}{\partial v}$. Right panels show the corresponding bifurcation diagrams.*

with an analogous system with an additional current I_{KS},

$$C\dot{v} = -\left[\bar{g}_{\text{L}}(v - E_{\text{L}}) + f_1(v, m) + \bar{g}_{\text{KS}}c(v - E_{\text{K}})\right]; \tag{3.22}$$

here $f_1(v, m) = \bar{g}_{\text{Ca}}n_\infty(v)(v - E_{\text{Ca}}) + \bar{g}_{\text{K}}m(v - E_{\text{K}})$ denotes the unchanged fast currents. Equations (3.21) and (3.22) are equivalent provided that we set the "leakage" factor c_{L} and the current I_{ext} in (3.21), respectively, equal to

$$c_{\text{L}} = 1 + \frac{\bar{g}_{\text{KS}}}{\bar{g}_{\text{L}}}\, c \quad \text{and} \quad I_{\text{ext}} = \bar{g}_{\text{KS}}\, c\, E_{\text{K}} + \bar{g}_{\text{L}}(1 - c_{\text{L}})E_{\text{L}}. \tag{3.23}$$

The desired bifurcation diagram of equilibrium voltage as a function of c is therefore a "slice" of the two-parameter $(v, \bar{g}_{\text{L}}c_{\text{L}})$ bifurcation surface above the line defined by eliminating c

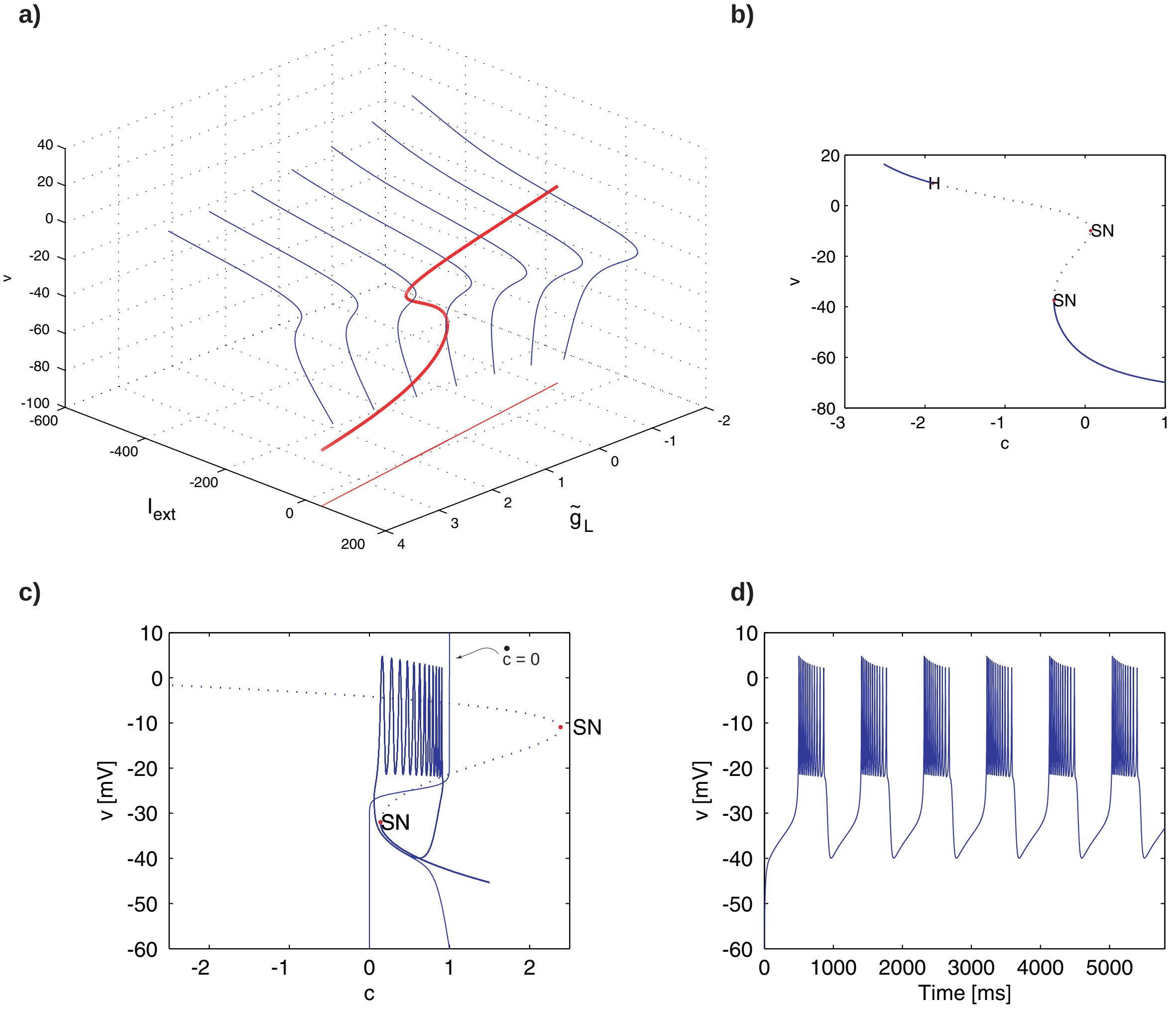


Figure 11. *Bifurcations with respect to c.* (a) *The two-parameter bifurcation surface as a function of I_{ext} and $\tilde{g}_L = \bar{g}_L c_L$; also shown is the bifurcation slice (bold) and its projection* (3.24) *(dashed).* (b) *The slice as a function of c; note the two SNs and an H bifurcation.* (c) *The $\dot{v} = 0$ and $\dot{c} = 0$ nullclines and a typical bursting trajectory projected onto the (c, v) plane.* (d) *Voltage time history exhibiting bursts.*

from (3.23):

$$I_{\mathrm{ext}} = \bar{g}_{\mathrm{L}}(E_{\mathrm{L}} - E_{\mathrm{K}})(1 - c_{\mathrm{L}}). \tag{3.24}$$

Figure 11 shows an example. Note that the line (3.24) is almost perpendicular to the I_{ext}-axis on Figure 11a; this is due to the fact that the difference between the leakage reversal potential E_{L} and the potassium reversal potential E_{K} is very small in this case; also note that the signs of the terms in (3.23)–(3.24) imply that the c-bifurcation diagram is reversed in comparison to the I-diagrams of Figures 8, 9, and 10, having the higher v branch extending to the left (see Figure 11b). Finally, we note that the maximal conductance associated with the very slow variable $\bar{g}_{\mathrm{KS}}$ does not influence the slice location (3.24). Rather, changes in $\bar{g}_{\mathrm{KS}}$, which enters the current I_{KS} of (3.22) multiplicatively, horizontally compress or expand the orbit projected on the (c, v) plane (Figure 11), a fact that will be useful in section 4.2.

3.3. The bursting mechanism. As anticipated at the beginning of this section, bursting results from hysteretic transitions between a quasi-static quiescent state and a periodic (spiking) state, driven by the slow variable c. Thus, given the bifurcation diagram in c, the dynamics of the third order model can be elucidated in the limit of small δ. Here we discuss a typical bifurcation diagram, with the sequence SN, SN, H as in Figure 11b (cf. Figure 10c$_2$ or Figure 8c$_2$). The vectorfield of (3.1c) indicates that $c \in [0, 1]$ will decrease when $c > c_\infty(v)$ and increase when $c < c_\infty(v)$. As c slowly evolves, the fast subsystem (3.1a)–(3.1b) remains close to its stable fixed point until the left-hand SN bifurcation on the lower branch is reached. When c passes this point, the state quickly jumps to the coexisting stable limit cycle (see Figure 11c). During this spiking oscillation, the average voltage is sufficiently high that c increases, until the cycle is destroyed as the limit cycle collides with the saddle point (the middle branch) in a saddle-loop (SL) or homoclinic bifurcation, or the right-hand SN occurs on the cycle itself (SNLC) [50]. Figure 11c shows the former case. It may also happen that the H bifurcation is subcritical [50] and the relevant stable limit cycle is born in an SN of periodic orbits (SNPO).

4. A minimal bursting model.

The bursting mechanism identified above includes a branch of stable equilibria terminating in an SN and a branch of limit cycles terminating in a global homoclinic bifurcation, or possibly destroyed by a second SN of fixed points occurring on the limit cycle. A minimal model therefore requires only the "nose" or NRC on the lower equilibrium branch, and an H bifurcation to create the periodic orbit on the upper branch. This can be captured by a fast persistent (inward) current. In the model discussed in [20], based on the two-variable Morris–Lécar equations [38], it is a calcium current; in Butera, Rinzel, and Smith's model 2 [17] (cf. Appendix C) it is a persistent sodium current I_{NaP}, with almost the same functional expression, the only difference being the exponent of the gating variable which is 1 in [38] and 3 in [17].

The following results were obtained for a persistent inward current with a reversal potential of $E = 120m\mathrm{V}$, consistent with calcium, which we called I_{Ca}. We believe that analogous results could be obtained with a persistent sodium current with reversal potential around $E = 50m\mathrm{V}$, but specific biophysical data is unavailable for CPG neurons in the cockroach, so we cannot identify a specific current, responsible for the fast spikes. In addition we have a slow (outward) current I_{K} and a leakage current I_{L}. The bursting mechanism will be caused by an additional very slow potassium current I_{KS} (essentially the same as I_{KS} in [17, Model 2]; see also [39]) that plays the same role of the calcium-activated potassium current I_{KCa} in the Sherman–Rinzel–Keizer (SRK) model: it hyperpolarizes (decreases) the membrane voltage when v is highly depolarized (i.e., in the bursting regime). Our main results should carry over when a calcium-dependent potassium current is used for the bursting mechanism as presented in the example below using the SRK model [13]. Therefore, we consider the system

$$\begin{aligned} C\dot{v} &= -[I_{\mathrm{Ca}} + I_{\mathrm{K}} + I_{\mathrm{L}} + I_{\mathrm{KS}}] + I_{\mathrm{ext}}, \\ \dot{m} &= \frac{\epsilon}{\tau_m(v)}\left[m_\infty(v) - m\right], \end{aligned} \tag{4.1}$$

$$\dot{c} = \frac{\delta}{\tau_c(v)} \left[c_\infty(v) - c\right].$$

The currents in (4.1) are specified by

$$\begin{aligned} I_{\mathrm{Ca}} &= \bar{g}_{\mathrm{Ca}} n_\infty(v)(v - E_{\mathrm{Ca}}), \qquad & I_{\mathrm{K}} &= \bar{g}_{\mathrm{K}} m \cdot (v - E_{\mathrm{K}}), \\ I_{\mathrm{L}} &= \bar{g}_{\mathrm{L}}(v - E_{\mathrm{K}}), & I_{\mathrm{KS}} &= \bar{g}_{\mathrm{KS}} c \cdot (v - E_{\mathrm{K}}). \end{aligned} \tag{4.2}$$

The steady state gating and timescale functions are of the types (2.3)–(2.4); in particular, $m_\infty(v)$ and $c_\infty(v)$ are both sigmoidal functions $\left(1 + e^{-k_0(v - v_{th})}\right)^{-1}$, where $m_\infty(v)$ is defined by $k_{0_{\mathrm{K}}}, v_{th_{\mathrm{K}}}$ and $c_\infty(v)$ by $k_{0_{\mathrm{KS}}}, v_{th_{\mathrm{KS}}}$. The parameters given in Table 1 were adopted for the work described in this section. The maximal conductances are expressed in $m\mathrm{S/cm}^2$, the reversal and threshold potentials in $m\mathrm{V}$, the slope coefficients in $m\mathrm{V/s}$, and the capacitance C in $\mu\mathrm{F/cm}^2$. All parameters excepting $C, \bar{g}_{\mathrm{K}}, \epsilon, \delta$ are the same as in Morris and Lécar [38, 20], $\bar{g}_{\mathrm{K}} = 9$ being slightly higher than their value $\bar{g}_{\mathrm{K}} = 8$. With the application to follow in [1] in mind, the parameters C, ϵ, and δ, which independently determine the time scales of v, m, and c, are set to match typical cockroach data.

Table 1

Parameter values for the bursting model.

$\bar{g}_{\mathrm{Ca}}$	=	4.4	E_{Ca}	=	120	$v_{th_{\mathrm{Ca}}}$	=	−1.2	$k_{0_{\mathrm{Ca}}}$	=	0.11
$\bar{g}_{\mathrm{K}}$	=	9.0	E_{K}	=	−80	$v_{th_{\mathrm{K}}}$	=	2.0	$k_{0_{\mathrm{K}}}$	=	0.2
$\bar{g}_{\mathrm{KS}}$	=	0.25				$v_{th_{\mathrm{KS}}}$	=	−27	$k_{0_{\mathrm{KS}}}$	=	0.8
$\bar{g}_{\mathrm{L}}$	=	2.0	E_{L}	=	−60						
C	=	1.2	ϵ	=	4.9	δ	=	0.052	I_{ext}	=	35.6

4.1. Silence, bursting, and beating. The existence of a resting potential and a limit cycle for the fast subsystem ensures that the cell can exhibit two states: silent or beating (persistent spiking). As we saw in section 3.3, to obtain bursting, these states must coexist over some parameter range. Moderate increases in external current I_{ext} leave the (v, c)-bifurcation diagram almost unchanged in shape but shift it rightward, causing the intersection of the nullclines to move from the lower, to the middle, and finally to the upper branch (Figures 12a$_1$–a$_4$). This effects a continuous change from silence to bursting to beating (Figures 12b$_1$–b$_4$). Similar results (not shown) can be obtained by changing the threshold voltage in the function $c_\infty(v)$.

The bursting frequency can be changed by over an order of magnitude (0.8–19.6 Hz) via the bias current I_{ext} (Figures 12(a$_2$,b$_2$)–(a$_3$,b$_3$)). This agrees with Butera, Rinzel, and Smith [17], in which variations from 0.05–1 Hz were found, but it is accompanied by an increase from five to nine APs. Since fast motoneurons encode force in terms of AP numbers, the latter should also be adjustable *without* substantial frequency change. This is possible in regimes with few APs per burst (Figures 12(b$_3$,b$_5$)).

4.2. Shaping the bursts. In the following we will concentrate on five parameters and show how they can affect the properties of the bursts. We anticipate that not all will be plausibly adjustable in vivo; in particular, we will show how one can fix the parameters C, ϵ, and δ to match timescales and key features in systems of interest, providing a "baseline" model, and how the adjustable parameters $(I_{\mathrm{ext}}, g_{\mathrm{KS}})$ affect this model. While I_{ext} cannot be

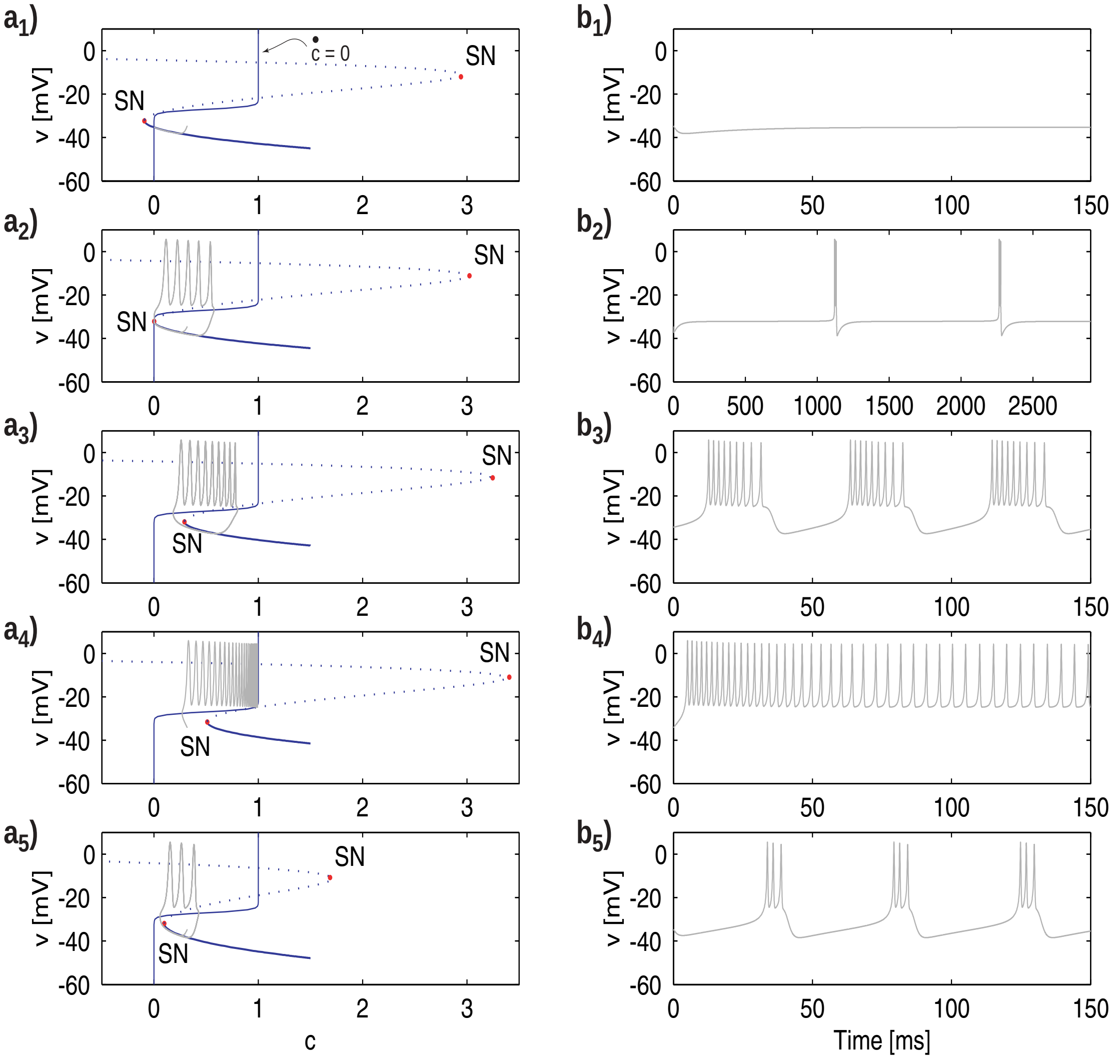


Figure 12. *Left panels* (a_1)–(a_5) *show bifurcation diagrams of the fast subsystem of* (4.1) *and projections of the bursting trajectory (grey line) onto* (c, v) *plane. The* $\dot{c} = 0$ *nullcline (solid) is also shown. Parameters are as in Table* 1*, unless stated otherwise. Right panels* (b_1)–(b_5) *show the membrane voltage* v *versus time.* (a_1)–(b_1) *Silence: the* $\dot{c} = 0$ *nullcline intersects the stable branch of the bifurcation diagram and there is one (stable) fixed point for* (4.1)*;* $I_{ext} = 34.5$. (a_2)–(b_2) *Low frequency bursting:* $f = 0.8$ *Hz;* $I_{ext} = 35.346$*. Notice that each burst has five spikes and note extended time scale in* (b_2). (a_3)–(b_3) *High frequency bursting:* $f = 19.6$ *Hz;* $I_{ext} = 38$. (a_4)–(b_4) *Beating: the system has a stable limit cycle with* $c \approx \text{const}$*;* $I_{ext} = 40$. (a_5)–(b_5) *Changing* $\bar{g}_{KS}$ *to* 0.35 *(in place of* $\bar{g}_{KS} = 0.19$ *in previous cases) contracts the bifurcation diagram, affecting the duty cycle;* $I_{ext} = 37$.

adjusted independently via, e.g., synapses from central nervous system (CNS) neurons, both it and $\bar{g}_{\mathrm{KS}}$ can be modulated by synaptic inputs and by neurotransmitters, so both of these control parameters are biophysically plausible in vivo.

(i) The capacitance C basically sets the frequency of the fast spiking. Here it was set to 1.2 in order to obtain fast spikes on the order of $1ms$.

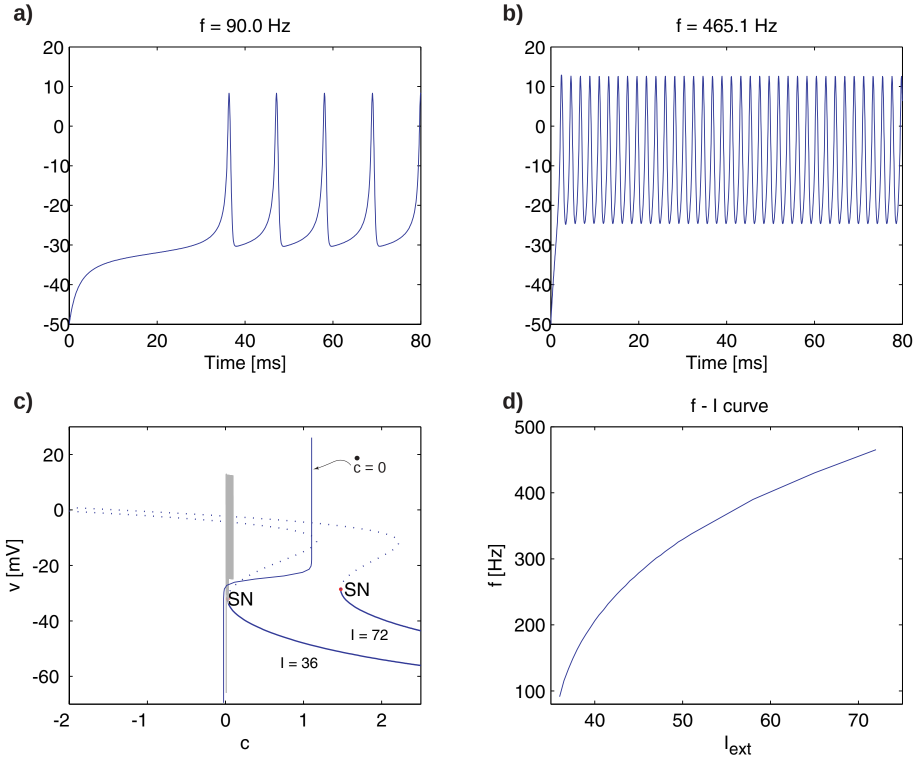


Figure 13. *Regulating spiking frequency by changing input current I_{ext}. Varying I_{ext} from* 36 (a) *to* 72 (b) *spans the frequency range from* 90 *to* 465 *Hz. The $\dot{v} = 0$ nullcline moves rightward as I_{ext} increases* (c)*, taking the limit cycle further from the homoclinic bifurcation.* (d) *The resulting $f - I$ curve. Parameters as in Table* 1*, except $\bar{g}_{KS} = 0.5$, $\epsilon = 2.0$, $\delta = 10^{-4}$.*

(ii) The parameter ϵ can play a central role as also suggested in the next illustrative example, at the end of section 4. Recalling Figure 5, we observe that the H condition $a_H = -\epsilon e C$ is the only one depending on the parameter ϵ. A decrease in ϵ therefore makes the curve $a_H = -\epsilon e C$ shallower, shifting the H bifurcation point to more depolarized levels. This, in turn, "drags" the global homoclinic bifurcation to the left. Assuming that the very slow dynamics is unchanged, this suggests that the number of spikes in a burst should decrease, because the homoclinic bifurcation moves closer to the SN bifurcation.

(iii) The parameter δ is responsible for the recovery variable time scale and therefore determines a "baseline" bursting frequency.

(iv) The bias current I_{ext} can have several effects. It can influence the bursting frequency, especially when the nullcline $\dot{c} = 0$ of the very slow variable is fairly close to the SN bifurcation (Figure 12(a_2)–(a_3)).

I_{ext} can affect the spiking frequency as shown in Figure 13, for which we set $\delta = 10^{-4}$ and bursting is so slow that behavior resembles a regular spiking neuron. (In the companion paper [1] this will be used to model slow cockroach motoneurons which spike in the range 90–400 Hz [52, 53].) A lower bound for maximum spiking frequency is given by $\omega_0 = \sqrt{\epsilon(ae+bd)/C}$ at the H bifurcation, and since the cycle is destroyed via a homoclinic bifurcation [50], there is no limit to minimum frequency in principle; however, away from the H bifurcation the $f-I$ curves for these neurons are rather flat (see, e.g., discussion in [41]: 53–138 Hz in the HH model, 19–28 Hz in FitzHugh–Nagumo, and 50–70 Hz in Morris–Lécar [38]). At least in the relaxation oscillator limit, frequency is essentially fixed by the dynamics on the slow manifold, which does not significantly change away from the H bifurcation (Figure 13d was obtained near the homoclinic bifurcation).

I_{ext} can also affect the number of APs per burst, but we note that the (percentage) variation is minimal when the number of APs per burst is large and becomes increasingly more important when there are few APs per burst ($\sim$ 4-5).

(v) The conductance g_{KS} is central in determining the duty cycle: the fraction of the period occupied by the burst. Recalling that the slice of the bifurcation diagram does not depend on $\bar{g}_{\text{KS}}$ (3.24), and that this maximal conductance enters multiplicatively in I_{KS} (3.22), we see that increases (decreases) in its value respectively expand (contract) the projected orbit in the c-direction, without changing the values of the corresponding $\bar{v}_{SN}$ and $\bar{v}_H$. The location of the homoclinic bifurcation responsible for disappearance of the cycle shifts in this deformation process. The time spent in each regime varies inversely with distance to the $\dot{c}=0$ nullcline; thus, in going from $\bar{g}_{\text{KS}} = 0.19$ to 0.35 (Figures 12a$_3$ to 12a$_5$), the quiescent fraction of the cycle increases since the lower branch of the $I_{ss}-v$ curve moves closer to the nullcline. Figures 14(a,b) show how bursting frequency and duty cycle can be independently changed by a suitable combination of the parameters I_{ext} and $\bar{g}_{\text{KS}}$. I_{ext} primarily affects frequency, especially at higher values of $\bar{g}_{\text{KS}}$; $\bar{g}_{\text{KS}}$ affects both frequency and duty cycle.

Summary. The model parameters C, ϵ, and δ in (4.1) may be chosen to match timescales of fast spikes (C), approximate number of APs per burst (ϵ), and baseline bursting frequency (δ). Depending on the number of APs per burst, two regimes can be identified: high ($\sim$ 15 APs) or low ($\sim$ 4 APs). In the high regime, bursting frequency is modulated by I_{ext}; in the low regime, I_{ext} influences both bursting frequency and number of APs per burst. In the high regime g_{KS} primarily affects the duty cycle; in the low regime it affects both duty cycle and number of APs per burst.

To satisfy changing behavioral demands CPGs must produce wide variations in cycle frequency, relative timing, and activity levels in motoneurons and muscles. One might therefore expect that the four key characteristics—bursting frequency, duty cycle, number of APs per burst, and spiking frequency—should be *independently* adjustable, since they serve different physiological functions (e.g., in locomotion, bursting sets the stepping frequency, and slow motoneuron spike rates and fast motoneuron APs determine muscle force, via calcium release dynamics). Such flexibility may seem impossible with only the two parameters I_{ext} and g_{KS}. Moreover, since conductance changes are slower, adjustments might not be possible on compatible timescales, and as we have noted above, I_{ext} is in any case not directly accessible in vivo.

Here we anticipate a solution that evolution may have achieved via “division of labor”; more details will be given in [1]. Insect CPGs comprise at least six bursting interneurons, each of which drives fast (bursting) motoneurons D$_f$ and slow (spiking) motoneurons D$_s$. Stepping

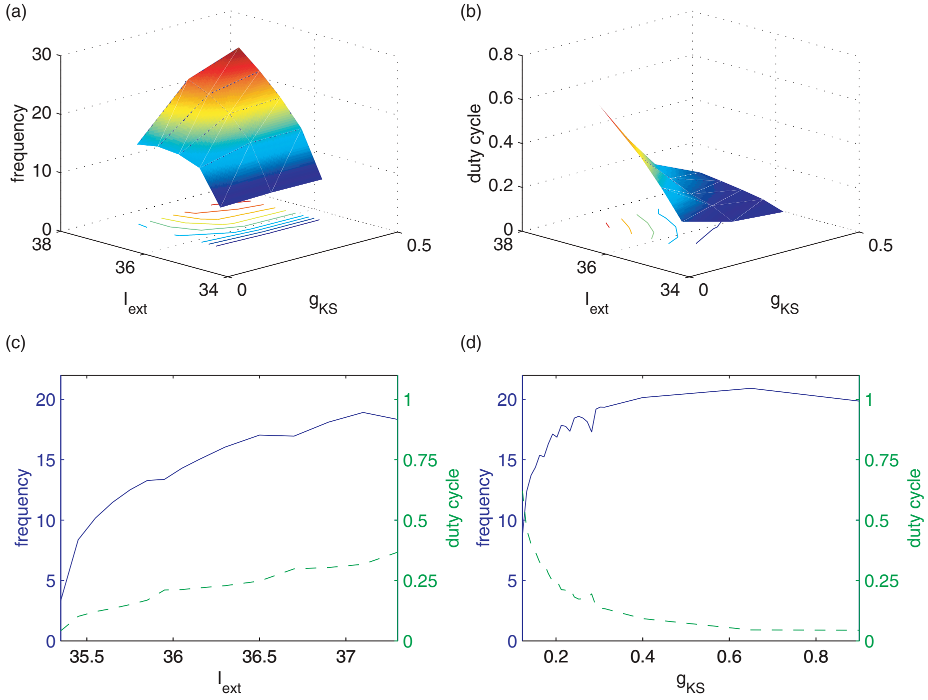


Figure 14. *Bursting frequency* (a) *and duty cycle* (b) *dependence on I_{ext} and $\bar{g}_{KS}$, showing that I_{ext} primarily affects frequency at higher values of $\bar{g}_{KS}$ (see contour lines), but $\bar{g}_{KS}$ simultaneously affects duty cycle and frequency.* (c), (d): *Slices for constant $\bar{g}_{KS} = 0.19$ and constant $I_{ext} = 36.5$; frequency shown solid and duty cycle dashed.*

frequency and duty cycle can be set at the network level by synaptic currents from CNS and local reflexive feedback circuits, which effectively change CPG input currents I_{ext} and conductances g_{KS}. For reasons to be explained in [1], these two parameters together with external currents to slow motoneurons completely define the operational regime of the latter. Fast motoneurons, which grade force via the number of APs per burst, require more subtle treatment. As noted above, they can be modulated by their input currents and conductances, but these parameters also affect their bursting frequencies. Here network properties come to the rescue: unilaterally connected motoneurons "follow" CPG neurons provided that their bursting frequencies are close enough, in which case they entrain to the CPG bursting frequency over finite current and conductance ranges. This renders the D_f bursting frequency independent of these parameters, which therefore affect only the number of APs per burst. In summary, independent controls can be obtained by synergy of individual and network properties using (i) three different sets of bursters with five biophysical parameters and (ii) a network with appropriate leader-follower connections.

4.3. An illustrative example. We now illustrate how the foregoing analysis can help one to modify existing models to produce desired behaviors, perhaps when precise parameter details, or even current types, are unavailable. Specifically, we show how the SRK model [13], first introduced by Chay and Keizer [12], can be adapted to yield different duty cycles and numbers of APs per burst.

The SRK model may be written

$$\begin{aligned} C_m\dot{v} &= -\bar{g}_{\mathrm{K}}n(v-E_{\mathrm{K}}) - I_{\mathrm{Ca}}(v) - g_{\mathrm{KCa}}(Ca)\cdot(v-E_{\mathrm{K}}), \\ \dot{n} &= \lambda\frac{n_\infty(v)-n}{\tau_n(v)}, \\ \dot{Ca} &= f(-\alpha I_{\mathrm{Ca}}(v) - k_{\mathrm{Ca}}Ca) \end{aligned} \tag{4.3}$$

(cf. [37] and a slightly modified version in [33, p. 192]). Here,

$$\begin{aligned} I_{\mathrm{Ca}}(v) &= \bar{g}_{\mathrm{Ca}}m_\infty(v)h_\infty(v)(v-E_{\mathrm{Ca}}), \\ g_{\mathrm{KCa}}(Ca) &= \bar{g}_{\mathrm{KCa}}\frac{Ca}{K_d+Ca}, \end{aligned} \tag{4.4}$$

and $n_\infty, m_\infty, h_\infty$ are standard HH-type equilibrium functions (see Appendix B for functional forms and parameters). The model has a potassium current $I_{\mathrm{K}} = \bar{g}_{\mathrm{K}}n(v-E_{\mathrm{K}})$, a fast transitory calcium current I_{Ca}, and a very slow calcium-dependent potassium current $I_{\mathrm{KCa}} = g_{\mathrm{KCa}}(Ca)\cdot(v-E_{\mathrm{K}})$. Intracellular calcium affecting the conductance via (4.4) has its own dynamics given in the last equation of (4.3).

To compare (4.3) with (4.1) more directly, we first rewrite the system so that the calcium-dependent potassium current $I_{\mathrm{KCa}} = g_{\mathrm{KCa}}(Ca)(v-E_{\mathrm{K}})$ is linear in a new very slow variable

$$c = \frac{Ca}{K_d+Ca}. \tag{4.5}$$

Differentiating (4.5), we find $\dot{c} = \frac{K_d}{(K_d+Ca)^2}\,\dot{Ca}$, and inverting (4.5) to obtain $Ca = K_d\,\frac{c}{1-c}$, we have

$$\begin{aligned} C_m\dot{v} &= -\bar{g}_{\mathrm{K}}n(v-E_{\mathrm{K}}) - I_{\mathrm{Ca}}(v) - \bar{g}_{\mathrm{KCa}}c(v-E_{\mathrm{K}}), \\ \dot{n} &= \lambda\frac{n_\infty(v)-n}{\tau_n(v)}, \\ \dot{c} &= f\frac{(1-c)^2}{K_d}\left(-\alpha I_{\mathrm{Ca}}(v) - k_{\mathrm{Ca}}K_d\,\frac{c}{1-c}\right), \end{aligned} \tag{4.6}$$

with I_{Ca} as given above. The nullclines of the $\dot{c}$ equation are now

$$c = 1 \quad\text{and}\quad c = \frac{\alpha\bar{g}_{\mathrm{Ca}}m_\infty(v)h_\infty(v)(v-E_{\mathrm{K}})}{\alpha\bar{g}_{\mathrm{Ca}}m_\infty(v)h_\infty(v)(v-E_{\mathrm{K}}) - k_{\mathrm{Ca}}K_d}.$$

For the parameters of [13] the nullcline $\dot{c} = 0$ behaves as in our model, in the relevant region of voltages (Figure 15a$_4$).

The current $\bar{g}_{\mathrm{KCa}}$ enters (4.3) as does $\bar{g}_{\mathrm{KS}}$ in (4.1)–(4.2); we can therefore expect that changing $\bar{g}_{\mathrm{KCa}}$ will primarily affect the duty cycle. Figures 15a$_1$–b$_1$ reveal that this is the case, although the bursting frequency also changes. In fact, since $K_d \gg Ca$, the conductance g_{KCa} is essentially proportional to c, (4.5) is in its linear regime, and (4.3)–(4.4) is close to (4.1)–(4.2).

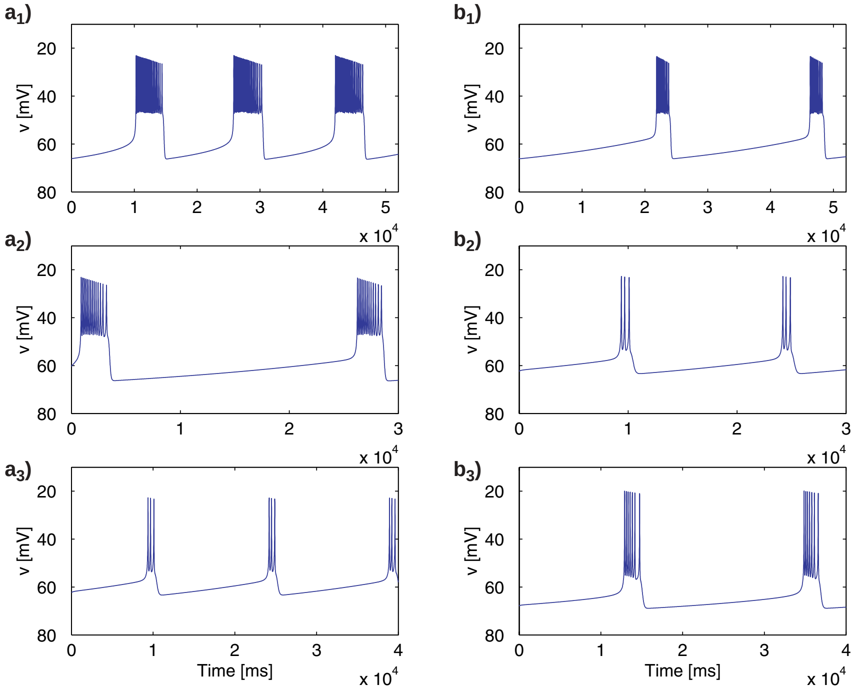


Figure 15. *Bursts in the SRK model. Panels* (a$_1$), (b$_1$) *Duty cycle changes due to* $\bar{g}_{KCa}$*:* $\bar{g}_{KCa} = 30000$*, left, and* 41750*, right.* (a$_2$), (b$_2$) *Effect of* λ *on numbers of APs:* $\lambda = 1.7$*, left, and* 1.55*, right.* (a$_3$), (b$_3$) *Effect of added external currents on AP numbers and bursting frequency:* $I_{ext} = 0$*, left, and* -550*, right.*

To adjust the number of spikes per burst, we could change membrane capacitance C, but this has very little effect on spike numbers and drastically reduces their magnitudes (results not shown). Adding bias currents also has little effect, since the system is in a high AP number/burst regime. However, decreasing the parameter λ ($\sim \epsilon$ in (4.1)) reduces the number of APs from 22 to 2–3; this is accompanied by a moderate increase in bursting frequency (Figure 15(a$_2$–b$_2$)). In this regime, an additional bias current has a much stronger influence, permitting adjustment of AP numbers without drastically changing the bursting frequency (Figure 15(a$_3$–b$_3$)).

5. Conclusions.

This paper develops a minimal model for a bursting neuron. We retain sufficient biophysical detail to permit appropriate parameter choices and variations to reproduce experimental data, while striving for generality and relative simplicity. Much current research concerns subcellular details of ionic currents, channels, and molecular messengers [43, 54, 17], but despite the ability of such detailed models to reproduce experimental data (e.g., [6, 7, 8]), their complexity and sensitivity to parameter variations render them effectively unanalyzable. We believe that massive simulations or experiments alone do not provide global understanding, which profits more from the identification of a few key mechanisms. We hope to extract these by judicious *selection*, rather than inclusion, of biological data, and in doing so to provide a flexible and tractable mathematical framework within which biological hypotheses can be investigated and novel experiments suggested.

To this end, we review ion channel models of HH type and propose a generic three-dimensional ODE (4.1)–(4.2) that exploits the presence of three disparate timescales and obviates the detailed analysis of multiple currents, although we show how additional currents can be classified and incorporated, and their influences investigated via steady state current-voltage curves and their derivatives. We note that some currents increase complexity without adding new qualitative behaviors, and thus can be neglected, at least in a first approximation. Our procedure yields guidelines for creating models of specific behaviors, and we use it here to select a minimal set of currents necessary to produce bursting, and to understand the role of biophysical parameters such as conductances and bias currents in determining the bursting frequency, duty cycle, spike rates, and numbers of APs per burst. We further illustrate by showing how duty cycles and AP numbers can be adjusted in the SRK model.

Previous work of Bertram et al. [21], Rinzel and Lee [18], Rinzel [19], Izhikevich [35], and others, summarized in [33], develops a topological classification of bursting mechanisms, based on the types of bifurcations that the fast subsystem undergoes as c (or I_{ext}) varies. This illuminates the phase space geometry. The present treatment is more analytical in nature and allows one to determine if specific currents with particular "influence windows" $U_i = [v_{th_i}, E_i]$ can introduce new folds and hence SN bifurcations, or otherwise change stability types of equilibria in the fast subsystem. Although our classification is in terms of steady state properties of ionic currents and does not reveal all details of the periodic orbits, it nonetheless allows one to adjust periodic orbit branches in the fast subsystem, via the reduced $\dot{c} = 0$ nullcline and $\dot{v} = 0$ bifurcation set, and hence to tune burst properties.

In the paper [1] we will show how the bursting model (4.1)–(4.2), along with a single equation describing synaptic dynamics, may be used as the basic subunit in building a model of an insect CPG and motoneurons.

Appendix A. A Rose–Hindmarsh model. Rose and Hindmarsh [14, p. 273] considered the following model for a repetitively firing neuron:

$$C\dot{v} = -\left[\bar{g}_{\text{Na}} m^3 h (v - E_{\text{Na}}) + g_{\text{L}}(v - E_{\text{L}}) + g_{\text{K}} n^4 (v - E_{\text{K}}) - g_{\text{A}} a^3 b (v - E_{\text{K}})\right] + I, \tag{A.1}$$

where the five gating variables m, h, n, a, b are described by the usual first order kinetics (2.2b). From this they obtained the third order system

$$\begin{aligned} C\dot{v} = &- \left[-3\bar{g}_{\text{Na}} m_\infty^3 q (v - E_{\text{Na}}) + 3A\bar{g}_{\text{Na}} b_\infty m_\infty^3 (v - E_{\text{Na}})\right] \\ &- [0.85\bar{g}_{\text{Na}} m_\infty (v - E_{\text{Na}}) + \bar{g}_{\text{L}}(v - E_{\text{L}}) + \bar{g}_{\text{K}} q (v - E_{\text{K}})] \\ &- [\bar{g}_{\text{s}} s_\infty (v - E_{\text{s}}) + \bar{g}_{\text{out}} z (v - E_{\text{K}}) - I], \\ \dot{q} = &\ \frac{q_\infty(v) - q}{\tau_q(v)}, \\ \dot{z} = &\ \frac{z_\infty(v) - z}{\tau_z(v)}. \end{aligned} \tag{A.2}$$

In reducing the six-dimensional model they employed a slow gating variable q that combines both sodium and potassium channels, and, numerically confiming that $\tau_b(v) \approx \tau_n(v)$, they replaced *both* τ_b *and* τ_n by the average value $\tau_q(v) = \frac{1}{2}(\tau_b(v) + \tau_n(v))$.

Appendix B. The SRK model. The SRK model results of section 4.3 were obtained for (4.3)–(4.4), where n_∞, m_∞, and h_∞ are the standard HH equilibrium functions

$$n_\infty = \frac{1}{1+e^{\frac{V_n - v}{S_n}}}, \quad m_\infty = \frac{1}{1+e^{\frac{V_m - v}{S_m}}}, \quad h_\infty = \frac{1}{1+e^{\frac{v - V_h}{S_h}}}, \tag{B.1}$$

and

$$\tau_n(v) = \frac{\gamma}{e^{\frac{v-\bar{V}}{a}} - e^{-\frac{v-\bar{V}}{b}}}, \quad \alpha = \frac{1}{2V_{\text{Cell}}F}. \tag{B.2}$$

The parameters used in section 4.3 are given in Table 2.

Table 2

Parameter values for the SRK model for bursting pancreatic β-cells.

$\bar{g}_{\text{Ca}}$	=	1400 pS	E_{Ca}	=	110 mV
$\bar{g}_{\text{K}}$	=	2500 pS	E_{K}	=	-75 mV
$\bar{g}_{\text{KCa}}$	=	30000 pS			
C_m	=	5310 fF	V_{Cell}	=	1150 μm^3
F	=	96.487 Coul/mMol	K_d	=	100 μMol
λ	=	1.7	k_{Ca}	=	0.03 m/s
V_n	=	-15 mV	S_m	=	5.6 mV
V_m	=	4 mV	S_m	=	14 mV
V_h	=	-10 mV	S_h	=	10 mV
a	=	65 mV	b	=	20 mV
γ	=	60 ms	$\bar{V}$	=	-75 mV
f	=	0.001			

Appendix C. Bursting pacemaker neurons in the pre-Bötzinger complex. Butera, Rinzel, and Smith [17] considered two possible models for bursting pacemaker neurons in the pre-Bötzinger complex. Model 1 takes the form

$$\begin{aligned} C\dot{v} &= -[I_{\text{NaP}} + I_{\text{Na}} + I_{\text{K}} + I_{\text{L}} + I_{\text{tonic-e}}] + I_{\text{app}}, \\ \dot{n} &= \frac{\epsilon}{\tau_n(v)}\left[n_\infty(v) - n\right], \\ \dot{h} &= \frac{\delta}{\tau_h(v)}\left[h_\infty(v) - h\right], \end{aligned} \tag{C.1}$$

with $I_{\text{tonic-e}}$ and I_{app} fixed biases and the other currents specified by

$$\begin{aligned} I_{\text{Na}} &= \bar{g}_{\text{Na}} m_\infty^3(v)(1-n)\cdot(v - E_{\text{Na}}), & I_{\text{K}} &= \bar{g}_{\text{K}} n^4 \cdot (v - E_{\text{K}}), \\ I_{\text{L}} &= \bar{g}_{\text{L}}(v - E_{\text{L}}), & I_{\text{NaP}} &= \bar{g}_{\text{NaP}} m_\infty(v) h \cdot (v - E_{\text{Na}}). \end{aligned} \tag{C.2}$$

The time course of inactivation of the sodium gating channel (h in the original HH equations [2]) as stated in [17] is "assumed to be of similar dynamics as n and is approximated by $h = (1-n)$" [55, 56]. As in Appendix A, this is an instance of a single gating variable (n) associated to two different ionic channels (I_{Na} and I_{K}).

Model 2 takes the form

$$\begin{aligned} C\dot{v} &= -[I_{\text{NaP}} + I_{\text{KS}} + I_{\text{Na}} + I_{\text{K}} + I_{\text{L}} + I_{\text{tonic-e}}] + I_{\text{app}}, \\ \dot{n} &= \frac{\epsilon}{\tau_n(v)}\left[n_\infty(v) - n\right], \\ \dot{k} &= \frac{\delta}{\tau_k(v)}\left[k_\infty(v) - k\right]. \end{aligned} \tag{C.3}$$

In addition to a leakage current $I_{\text{L}} = \bar{g}_{\text{L}}(v - E_{\text{L}})$, the currents in (C.3) are

$$\begin{array}{lclclcl} I_{\text{Na}} & = & \bar{g}_{\text{Na}} m_\infty^3(v)\cdot(v - E_{\text{Na}}), & \qquad & I_{\text{K}} & = & \bar{g}_{\text{K}} n^4 \cdot (v - E_{\text{K}}), \\ I_{\text{KS}} & = & \bar{g}_{\text{KS}} k(v - E_{\text{K}}), & & I_{\text{NaP}} & = & \bar{g}_{\text{NaP}} m_\infty(v)\cdot(v - E_{\text{Na}}). \end{array} \tag{C.4}$$

Note that I_{NaP} does not inactivate as in model 1.

Acknowledgment. We thank the referees for their close reading and for many valuable suggestions.

REFERENCES

[1] R. Ghigliazza and P. Holmes, *A minimal model of a central pattern generator and motoneurons for insect locomotion*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 671–700.

[2] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerves*, J. Physiology, 117 (1952), pp. 500–544.

[3] L. F. Abbott, *Single neuron dynamics*, in Neural Modeling and Neural Networks, F. Ventriglia, ed., Pergamon Press, Oxford, UK, 1994, pp. 57–78.

[4] H. Fischer, J. Schmidt, R. Haas, and A. Büschges, *Pattern generation for walking and searching movements of a stick insect leg.* I. *Coordination of motor activity*, J. Neurophysiology, 85 (2001), pp. 341–353.

[5] J. Schmidt, H. Fischer, and A. Büschges, *Pattern generation for walking and searching movements of a stick insect leg.* II. *Control of motoneuronal activity*, J. Neurophysiology, 85 (2001), pp. 354–361.

[6] S. Grillner, P. Wallén, L. Brodin, and A. Lansner, *Neuronal network generating locomotor behavior in lamprey*, Annual Review in Neuroscience, 14 (1991), pp. 169–199.

[7] S. Grillner, *Bridging the gap—from ion channels to networks and behaviour*, Current Opinion in Neurobiology, 9 (1999), pp. 663–669.

[8] S. Grillner and P. Wallén, *Cellular basis of a vertebrate locomotor system—steering, intersegmental and segmental co-ordination and sensory control*, Brain Research Review, 40 (2002), pp. 92–106.

[9] A. H. Cohen, P. Holmes, and R. H. Rand, *The nature of coupling between segmental oscillators of the lamprey spinal generator for locomotion: A model*, J. Math. Biol., 13 (1982), pp. 345–369.

[10] N. Kopell, *Toward a theory of modelling generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. Cohen, S. Rossignol, and S. Grillner, eds., Wiley, New York, 1988, pp. 369–413.

[11] T. L. Williams, *Phase coupling by synaptic spread in chains of coupled neuronal oscillators*, Science, 258 (1992), pp. 662–665.

[12] T. R. Chay and J. Keizer, *Minimal model for membrane oscillations in the pancreatic β-cell*, Biophysical J., 42 (1983), pp. 181–190.

[13] A. Sherman, J. Rinzel, and J. Keizer, *Emergence of organized bursting in clusters of pancreatic β-cells by channel sharing*, Biophysics J., 54 (1988), pp. 411–425.

[14] R. M. Rose and J. L. Hindmarsh, *The assembly of ionic currents in a thalamic neuron* I. *The three-dimensional model*, Proc Roy. Soc. London Ser. B Biol. Sci., 237 (1989), pp. 267–288.

[15] R. M. Rose and J. L. Hindmarsh, *The assembly of ionic currents in a thalamic neuron* II. *The stability and state diagrams*, Proc Roy. Soc. London Ser. B Biol. Sci., 237 (1989), pp. 289–312.

[16] R. M. Rose and J. L. Hindmarsh, *The assembly of ionic currents in a thalamic neuron* III. *The seven-dimensional model*, Proc Roy. Soc. London Ser. B Biol. Sci., 237 (1989), pp. 313–334.

[17] R. J. Butera, Jr., J. Rinzel, and J. C. Smith, *Models of respiratory rhythm generation in the pre-Bötzinger complex.* I. *Bursting pacemaker neurons*, J. Neurophysiology, 81 (1999), pp. 382–397.

[18] J. Rinzel and Y. S. Lee, *On different mechanisms for membrane potential bursting*, in Nonlinear Oscillations in Biology and Chemistry, H. G. Othmer, ed., Lecture Notes in Biomath. 66, Springer-Verlag, Berlin, 1986, pp. 19–33.

[19] J. Rinzel, *A formal classification of bursting mechanisms in excitable systems*, in Mathematical Topics in Population Biology, Morphogenesis, and Neurosciences, E. Teramoto and M. Yamaguti, eds., Lecture Notes in Biomath. 71, Springer-Verlag, Berlin, 1987, pp. 267–281.

[20] J. Rinzel and G. B. Ermentrout, *Analysis of excitability and oscillations*, in Methods in Neuronal Modeling: From Synapses to Networks, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1989, pp. 135–169.

[21] R. Bertram, M. J. Butte, T. Kiemel, and A. Sherman, *Topological and penomenological classification of bursting oscillations*, Bull. Math. Biol., 57 (1980), pp. 413–439.

[22] X. J. Wang and J. Rinzel, *Oscillatory and bursting properties of neurons*, in Brain Theory and Neural Networks, M. A. Arbib, ed., MIT Press, Cambridge, MA, 1995, pp. 686–691.

[23] J. L. Hindmarsh and R. M. Rose, *A model of neuronal bursting using three coupled first order differential equations*, Philosophical Transactions of the Royal Society B: Biological Sciences, 221 (1984), pp. 87–102.

[24] M. Pernarowski, *Fast subsystem bifurcations in a slowly varying Liénard system exhibiting bursting*, SIAM J. Appl. Math., 54 (1994), pp. 814–832.

[25] G. de Vries, *Multiple bifurcations in a polynomial model of bursting oscillations*, J. Nonlinear Sci., 8 (1998), pp. 281–316.

[26] R. J. Butera, Jr., J. Rinzel, and J. C. Smith, *Models of respiratory rhythm generation in the pre-Bötzinger complex.* II. *Populations of coupled pacemaker neurons*, J. Neurophysiology, 81 (1999), pp. 398–415.

[27] J. Rinzel, *Discussion: Electrical excitability of cells, theory and experiment: Review of the Hodgkin-Huxley foundation and update*, Bull. Math. Biol., 52 (1990), pp. 5–23.

[28] J. Simmers, P. Meyrand, and M. Moulins, *Modulation and dynamic specification of motor rhythm-generating circuits in crustacea*, J. Physiology (Paris), 89 (1995), pp. 195–208.

[29] R. M. Rose and J. L. Hindmarsh, *A model of a thalamic neuron*, Proc. Roy. Soc. London Ser. B Biol. Sci., 225 (1985), pp. 161–193.

[30] R. Fitzhugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophysical Journal, 1 (1961), pp. 445–466.

[31] J. L. Hindmarsh and R. M. Rose, *A model for rebound bursting in mammalian neurons*, Philosophical Transactions of the Royal Society B: Biological Sciences, 346 (1994), pp. 129–150.

[32] P. Smolen, D. Terman, and J. Rinzel, *Properties of a bursting model with two slow inhibitory variables*, SIAM J. Appl. Math., 53 (1993), pp. 861–892.

[33] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.

[34] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Appl. Math. Sci. 126, Springer-Verlag, New York, 1997.

[35] E. M. Izhikevich, *Neural excitability, spiking and bursting*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 10 (1999), pp. 1171–1266.

[36] C. K. R. T. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, Lecture Notes in Math. 1609, Springer-Verlag, Berlin, 1995, pp. 44–118.

[37] J. Rinzel and G. B. Ermentrout, *Analysis of excitability and oscillations*, in Methods in Neuronal Modeling: From Ions to Networks, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1999, pp. 251–291.

[38] C. Morris and H. Lécar, *Voltage oscillations in the barnacle giant muscle*, Biophysics Journal, 35 (1981), pp. 193–213.

[39] S. D. Johnston and M.-S. Wu, *Foundations of Cellular Neurophysiology*, MIT Press, Cambridge, MA, 1995.

[40] G. W. Beeler and H. J. Reuter, *Reconstruction of the action potential of ventricular myocardial fibers*, J. Physiology, 268 (1977), pp. 177–210.

[41] C. Koch, *Biophysics of Computation*, Oxford University Press, New York, 1999.

[42] R. E. Plant, *Bifurcation and resonance in a model for bursting nerve cells*, J. Math. Biol., 11 (1981), pp. 15–32.

[43] J. Keizer and P. Smolen, *Bursting electrical activity in pancreatic β cells caused by Ca^{2+}- and voltage-inactivated Ca^{2+} channels*, Proc. Natl. Acad. Sci. USA, 88 (1991), pp. 3897–3901.

[44] P. Dayan and L. Abbott, *Theoretical Neuroscience*, MIT Press, Cambridge, MA, 2001.

[45] J. A. Connor and C. F. Stevens, *Inward and delayed outward membrane currents in isolated neural somata under voltage clamp*, J. Physiology, 213 (1971), pp. 1–19.

[46] D. A. McCormick and J. R. Huguenard, *A model of the electrophysiological properties of thalamocortical relay neurons*, J. Neurophysiology, 68 (1992), pp. 1384–1400.

[47] R. Ghigliazza, *Neuromechanical Models for Insect Locomotion*, Ph.D. thesis, Princeton University, Princeton, NJ, 2004.

[48] W. A. Wilson and H. Wachtel, *Negative resistance characteristic essential for the maintenance of slow oscillations in bursting neurons*, Science, 186 (1974), pp. 932–934.

[49] J. Rinzel and Y. S. Lee, *Dissection of a model for neuronal parabolic bursting*, J. Math. Biol., 25 (1987), pp. 653–675.

[50] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983; corrected sixth printing, 2002.

[51] J. A. Connor and C. F. Stevens, *Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma*, J. Physiology, 213 (1971), pp. 31–53.

[52] K. G. Pearson, *Central programming and reflex control of walking in the cockroach*, J. Experimental Biology, 56 (1972), pp. 173–193.

[53] A. K. Tryba and R. E. Ritzmann, *Multi-joint coordination during walking and foothold searching in the Blaberus cockroach.* II. *Extensor motor patterns*, J. Neurophysiology, 83 (2000), pp. 3337–3350.

[54] P. Varona, J. J. Torres, R. Huerta, H. D. I. Abarbanel, and M. I. Rabinovich, *Regularization mechanisms of spiking-bursting neurons*, Neural Networks, 14 (2001), pp. 865–875.

[55] V. I. Krinskii and Y. M. Kokoz, *Analysis of equations of excitable membranes* I. *Reduction of the Hodgkin-Huxley equations to a second-order system*, Biofizika, 18 (1973), pp. 506–511.

[56] J. Rinzel, *Excitation dynamics: Insights from simplified membrane models*, Fed. Proc., 44 (1985), pp. 2944–2946.

SIAM J. Applied Dynamical Systems
Vol. 3, No. 4, pp. 671–700

A Minimal Model of a Central Pattern Generator and Motoneurons for Insect Locomotion[*]

R. M. Ghigliazza[†] and P. Holmes[‡]

Abstract. We adapt the generic three-dimensional bursting neuron model derived in the companion paper [*SIAM J. Appl. Dyn. Syst.*, 3 (2004), pp. 636–670] to model central pattern generator interneurons and slow and fast motoneurons in insect locomotory systems. Focusing on cockroach data, we construct a coupled network that retains sufficient detail to allow investigation and prediction of biophysical parameter changes. We show that the model can encompass stepping frequency, duty cycle, and motoneuron output variations observed in cockroaches, and we reduce it to an analytically tractable symmetric network of coupled phase oscillators from which general principles can be extracted. The model's modular form allows dynamical analyses of individual components and the addition of other components, so we expect it to be more generally useful.

Key words. central pattern generators, bursting neurons, motoneurons, insect gaits, phase response curves, averaging, bifurcation, stability

AMS subject classifications. 34C15, 34C25, 34C29, 37Gxx, 92B05, 92B20, 92C20

DOI. 10.1137/040607563

1. Introduction.

Central pattern generators (CPGs) are networks of functionally distinguishable neurons, located in the vertebrate spinal cord or in invertebrate thoracic ganglia, capable of generating and regulating the spatio-temporal activity of motoneurons in the absence of sensory input (e.g., [2, 3, 4]). Over forty years of in vitro and in vivo studies of network architectures, intrinsic membrane properties, and neuromodulators (e.g., [5, 6, 7, 4, 8]) have firmly established their importance in motor behavior. CPG dynamics depends on intracellular, synaptic, and network level phenomena and can display remarkable richness and flexibility.

In this paper, using the reduced bursting neuron ODEs derived and studied in the preceding paper [1], we develop a model of the CPG and associated bursting motoneurons for insect locomotion. We draw on data from the death's head and American cockroaches *Blaberus discoidalis* and *Periplaneta americana* and focus on rapid running, a regime in which preflexive feedforward control [9, 10] appears to dominate and reflexive feedback plays a less important role [11, 12, 13] than in, e.g., stick insects [14] that use more varied gaits and leg placement strategies. We include enough ionic current and conductance detail to reveal how modulation

[*] Received by the editors December 8, 2003; accepted for publication (in revised form) by M. Golubitsky July 2, 2004; published electronically December 27, 2004. This work was supported by DoE, DE-FG02-95ER25238.

http://www.siam.org/journals/siads/3-4/60756.html

[†] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544 (rghiglia@princeton.edu). The work of this author was partially supported by a Burroughs-Wellcome Training Grant in Biological Dynamics, 1001782, and a Britt and Eli Harari Fellowship from the Department of Mechanical and Aerospace Engineering, Princeton University.

[‡] Department of Mechanical and Aerospace Engineering and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544 (pholmes@math.princeton.edu).

of specific biophysical parameters can adjust CPG outputs that determine key locomotive properties, while providing sufficient tractability to enable mathematical analysis. In particular, after reduction of the CPG "core" to phase equations, simple symmetry arguments locate fixed points and describe gaits, and eigenvalue calculations determine their stability properties. We also confirm that these phase reductions correctly represent the dynamics of the full network.

While we emphasize the double-tripod gait employed by *Blaberus* over the speed range 10–60 cm sec^{-1} [15, 16] (and common to many insects), our model also describes other gait patterns, and its modularity will permit the inclusion of additional inter- and motoneurons and reflexive sensing. Indeed, our ultimate goal is to marry it to muscle and body-limb mechanical models of the types developed in [17, 18, 19, 20] and to equip the whole with proprioceptive feedback and goal-oriented direction.

This paper is organized as follows. In section 2 we recall the bursting neuron model of [1] and summarize the effects of bias currents and conductances on its behavior. We then review relevant data on CPG neurons, motoneurons, and network architectures in section 3 and use it to assemble a hexapedal pattern generator in section 4. This comprises six synaptically interconnected CPG bursters, each driving a fast and a slow bursting motoneuron. Finally, to permit elementary analyses of network properties, in section 5 we reduce, via phase response curves (PRCs) and averaging, to phase variables alone. We summarize and outline future work in section 6.

2. A minimal bursting model.

The analyses developed in the preceding paper [1] enable us to propose a model sufficiently general to apply to both bursting CPG neurons and motoneurons. It includes a branch of stable equilibria terminating in a saddle-node and one of limit cycles terminating in a global homoclinic bifurcation, separated by a branch of unstable (saddle-type) equilibria. (Index theory [21] implies that the periodic orbits must encircle unstable equilibria, so an upper equilibrium branch also must exist.) A minimal model requires only the saddle-node on the lower equilibrium branch and a Hopf bifurcation to create the periodic orbit on the upper branch. As shown in [1], this can be captured by a fast nonlinear current, e.g., I_{Ca}, a leakage current I_{L}, a slow potassium current I_{K}, and an additional very slow current, I_{KS}, giving the system [1, equation (31)]

$$\begin{aligned} C\dot{v} &= -[I_{\mathrm{Ca}} + I_{\mathrm{K}} + I_{\mathrm{L}} + I_{\mathrm{KS}}] + I_{\mathrm{ext}}, \\ \dot{m} &= \frac{\epsilon}{\tau_m(v)}\left[m_\infty(v) - m\right], \\ \dot{c} &= \frac{\delta}{\tau_c(v)}\left[c_\infty(v) - c\right]. \end{aligned} \tag{2.1}$$

The currents appearing in (2.1) are

$$\begin{aligned} I_{\mathrm{Ca}} &= \bar{g}_{\mathrm{Ca}} n_\infty(v)(v - E_{\mathrm{Ca}}), \qquad & I_{\mathrm{K}} &= \bar{g}_{\mathrm{K}} m \cdot (v - E_{\mathrm{K}}), \\ I_{\mathrm{L}} &= \bar{g}_{\mathrm{L}}(v - E_{\mathrm{K}}), & I_{\mathrm{KS}} &= \bar{g}_{\mathrm{KS}} c \cdot (v - E_{\mathrm{K}}), \end{aligned} \tag{2.2}$$

where the steady state gating variables $m_\infty(v), n_\infty(v), c_\infty(v)$ and time "constants" $\tau_m(v), \tau_c(v)$

take the forms

$$w_{i_\infty}(v; k_{i_0}, v_{i_{th}}) = \frac{1}{1 + e^{-k_{i_0}(v - v_{i_{th}})}}, \tag{2.3}$$

$$\tau_i(v; k_{i_0}, v_{i_{th}}) = \operatorname{sech}\left(k_{i_0}(v - v_{i_{th}})\right), \tag{2.4}$$

with $w_{i_\infty} = m_\infty(v), n_\infty(v), c_\infty(v)$. Parameters were generally fixed as specified in Table 1 of section 4; modifications will subsequently be made to accomodate other behaviors. All parameters excepting $C, \bar{g}_\text{K}, \epsilon, \delta$ are the same as in Morris and Lécar [22, 23], $\bar{g}_\text{K} = 9$ being slightly higher than their value $\bar{g}_\text{K} = 8$. The parameters C, ϵ, and δ, which independently determine the time scales of v, m, and c, are set to match typical cockroach data.

As shown in section 4 of [1], variations in three key characteristics of the bursting pattern can be achieved quasi-independently by varying two biophysical parameters for each of the different neuron types to be used in the model. Specifically, in the appropriate regimes, the following hold:

1. The bursting frequency can be adjusted primarily by I_ext.
2. The spiking frequency can be adjusted by I_ext.
3. The number of action potentials (APs) can be adjusted by I_ext and $\bar{g}_\text{KS}$.
4. The duty cycle can be adjusted by $\bar{g}_\text{KS}$, although this may also affect frequency.

As we shall see, the bursting frequency and duty cycles of CPG interneurons are primarily responsible for speed adjustment (although leg extension, via stride lengths, is also important at higher speeds [24]), while motoneuron spiking frequencies and AP numbers grade force production. We remark that I_ext can be modulated by excitatory and inhibitory synapses from CNS neurons, and $\bar{g}_\text{KS}$ by suitable neurotransmitters, so both of these are biophysically plausible control parameters in vivo.

3. CPG neurons and motoneurons as bursters.

Before proposing specific parameter regimes for cockroach CPG and motoneuron models, we review relevant data on animals and insects in general and cockroaches in particular. We start by briefly commenting on spiking and nonspiking interneurons in CPGs, turn to motoneurons, and then discuss network connectivity.

3.1. CPG neurons: Bursters and nonspikers.

Working from direct recordings and deafferented (sensorless) preparations of the American cockroach *Periplaneta americana* [25, 26], Pearson [27, 28] hypothesized a flexor burst generator for each leg that comprises several interneurons, including a bursting interneuron that periodically excites the flexor (levator or swing) motoneurons while inhibiting the extensor (depressor or stance) units. Subsequently, interneurons that do not produce APs were found [29, 28, 30], and their importance in generating motor patterns was stressed. (In the locust they are responsible for coordinating subsets of motoneurons, controlling their spiking frequencies, and altering reflex strengths and movement magnitudes in a continuous and precise manner through graded potentials [31, 32].) Despite this, there is no evidence that nonspiking interneurons exhibit pacemaking capabilities; indeed, their quasi-sinusoidal membrane voltages could simply result from integration of incoming bursts [33]. Thus, while they may be involved in CPG circuits and may contribute in a graded manner to slow motoneuron outputs, we shall omit them from our model.

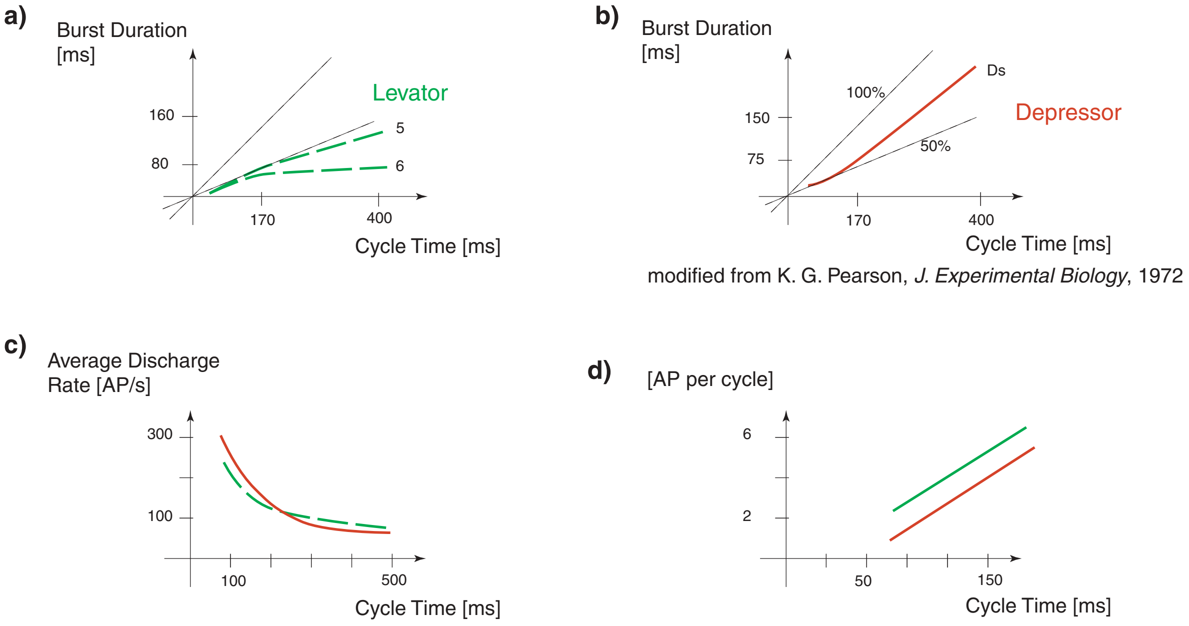
Plateau potentials, slow voltage oscillations on which the fast spikes ride, do however seem crucial to bursting [34]. These derive from bistability of the type illustrated in the bifurcation diagrams of Figures 9 and 10 of [1] for the fast subsystem, which allows brief inputs to trigger activities that outlast input duration; similarly, brief inhibitory stimuli can terminate plateaus [35]. This nonlinear membrane property plays a pivotal role in structuring bursts and producing cyclic behavior with appropriate time scales for stepping frequencies; it also allows brief proprioceptive inputs to reset and regulate the rhythm. We shall therefore represent each of the six "leg units" of an insect CPG by a single bursting (inter-) neuron of the form (2.1)–(2.2), synapsing directly on motoneurons innervating the dominant depressor muscles. Our model allows for subsequent addition of nonspiking interneurons between CPG and motoneurons.

3.2. Motoneurons. Because of important constraints imposed by their physiology, we discuss motoneurons and muscles in some detail.

The basic functional component of motor pathways is the motor unit, consisting of a motoneuron and the muscle fibers innervated by it. A single AP in the motoneuron causes a contractive twitch in the muscle fibers to which it is attached. Three types of motor units can be distinguished by their motoneuron firing patterns and muscle fiber properties. Slow twitch (S-type) units take about 50 msec to develop peak force and show little decline in force over prolonged periods of repetitive stimulation; they can exert low forces for very long periods. In contrast, fatigue resistant (FR) and fast fatigue (FF) units maximally contract in 5–10 msec. With repetitive stimuli, FR units can sustain moderate forces for ≈ 5 min before steady decline sets in over many minutes. FF motor units can achieve the greatest force of the three types, but with repetitive stimuli the force drops precipitously after 30 seconds or so. Both FR and FF units produce rapid large forces and so are found preferentially in muscles involved in executing fast movements. In the cockroach, slow and fast motoneuron discharges are quite distinct; slow units spike continuously at rates from 100–400 Hz when active [26, 36], while fast units typically produce 1–6 large spikes during a 50–100 msec stance or swing phase [37].

Muscle contraction force is determined by the motor pool in two ways. Small force increases are primarily met by greater motoneuron firing rates, but for larger contractions the number of active motoneurons is increased in a process called recruitment. This occurs in an orderly manner in the sequence S-FR-FF [38], determined jointly by the effect of cell body size [39] on excitatory postsynaptic potentials and on graded inputs to S, FR, and FF units. An incoming (tonic) stimulus sequentially excites the units as it passes their different thresholds [40, 34]. Cockroach coxal depressor motoneurons are innervated by both fast and slow motoneurons [41], and in *Blaberus* fast motoneuron recruitment begins at leg cycle rates of ≈ 6 Hz, corresponding to running speeds of 12 cm sec^{-1} [37]; fast motoneurons dominate at high speeds, but there is a considerable "overlap range" [26].

Often only slow motoneurons are modeled. Since these exhibit continuous relationships between firing frequency and force production, compact reductions of the whole neuromotor complex are then possible (e.g., the neuromuscular transforms of [42]). However, since we focus on rapid running, in which fast motoneurons are involved, and there is strong evidence of plateau and bursting capabilities in cockroach motoneurons [35], spiking and bursting



modified from K. G. Pearson, *J. Experimental Biology*, 1972

modified from K. G. Pearson, *J. Experimental Biology*, 1972

modified from R. J. Full, *J. Experimental Biology*, 1998

Figure 1. (a) *Burst duration in levator (dashed, axons* 5 *and* 6*) and* (b) *slow depressor* D_s *axon (solid) as a function of cycle time. The two thin dashed lines indicate duty cycles of* 100% *and* 50%. *Note that duty cycles of both depressors and levators approach* 50% *as speed increases (cycle time decreases). Cycle frequency ranges from* 2 *to* 10 *Hz.* (c) *Average spike rate of levator (dashed) and depressor (solid) axons. From Pearson* [26]. (d) *Approximate numbers of muscle action potentials (MAPs) per cycle in metathoracic (upper curve) and mesothoracic muscle (lower curve): Regression equations* $MAPs = 0.051t_{cyc} - 2.5$, $R^2 = 0.52$ *for metathoracic and* $MAPs = 0.048t_{cyc} - 3.2$, $R^2 = 0.42$ *for mesothoracic. From Full et al.* [37].

behaviors cannot be ignored. Indeed, given the few large spikes typically seen during rapid running, spike times and interspike intervals may be crucial in determining relative forces in different legs of the stance tripod and in regulating episodes of negative and positive work [37]. These aspects are certainly as important as the analogous role of slow motoneuron spiking frequency in low-speed walking. For this reason, and to allow continuous transition from slow to fast speeds, we shall use the bursting model (2.1), with suitable parameter choices, to represent both fast and slow motoneurons.

In Figures 1(a)–(c) we reproduce *Periplaneta* data from [26, 27] showing burst durations and spiking rates of slow cockroach motoneurons as functions of cycle time or inverse stepping frequency (cf. [43, 36] for analogous and more recent *Blaberus* data). In Figure 1(d) we reproduce data from [37] showing the dependence of number of APs in fast motoneurons as a function of cycle time. Phase relationships among leg muscles (not shown) indicate near constant antiphase between motoneurons associated with the left and right tripods. This data will guide our parameter choices.

3.3. Network configuration. Apart from anatomic identification and the acceptance of some degree of hierarchy [44, 45, 34], the precise division of labor among the higher central nervous system (CNS), the CPG-motoneuron complex, and proprioceptive sensing and

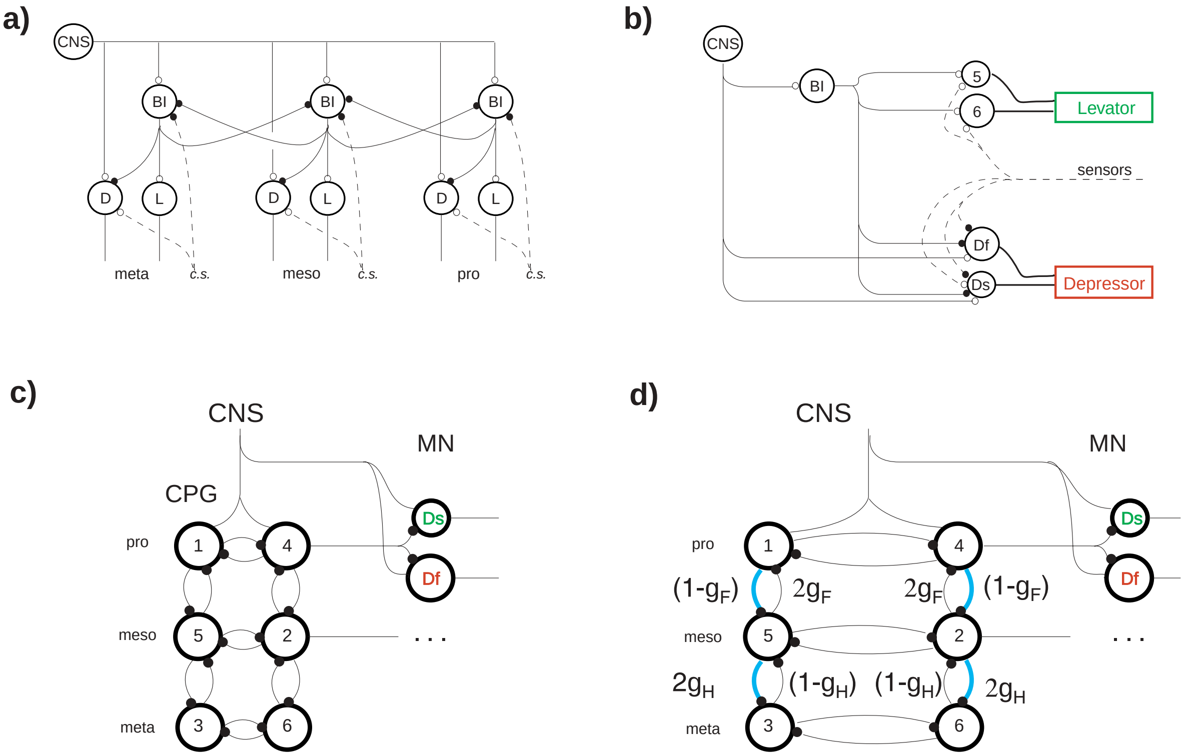


Figure 2. (a) *Ipsilateral CPG-motoneuron network connectivity and* (b) *individual leg depressor and levator circuit showing fast and slow motoneurons D_f, D_s as proposed by Pearson* [27]*. The CNS excites the bursting interneurons (BI) as well as D_f and D_s, which innervate the depressor muscles. Motoneurons* 5 *and* 6 *innervate the levator muscles and are not modeled here. Sensory feedback (dashed) affects the activity of all motoneurons. Open circles indicate excitatory coupling, and closed circles indicate inhibitory coupling.* (c) *Network connectivity of the hexapedal model. CPG neurons are coupled through mutually inhibiting synapses, and fast and slow motoneurons are connected via an inhibitory synapse to their corresponding CPG neuron; they are also tonically driven by the CNS. Sensory feedback is briefly discussed in the text but is not explicitly modeled. For synaptic weights, see text.* (d) *Asymmetric coupling as considered in section* 5.6*; the network of* (c) *is obtained with $g_F = \frac{1}{2} = g_H$.*

feedback remains unclear, but following Wilson [5], Pearson and Iles [27] deduced some general principles from the experiments noted in section 3.1. They found evidence of mutual inhibition between CPG interneurons belonging to the motor complexes of neighboring ipsilateral legs; they also found that CPG (inter)neurons excite levator motoneurons (active during the swing phase) but *inhibit* depressor motoneurons (active during stance), that sensory signals from campaniform sensillae and hair plates [46] (force and positions in the legs) tend to excite depressor motoneurons, and that tonic CNS signals generally excite both CPG neurons and depressor motoneurons. Their proposed architecture is reproduced in Figures 2(a), (b). The subsequent discovery of an interneuron (the lambda cell) involved in the escape response and highly depolarized by inputs from the ipsilateral campaniform sensilla and the contralateral trochanteral hair plate [46] supports this picture.

Henceforth we exclude levators, since our mechanical models neglect leg masses and the swing phase is implicit [17, 12, 20]. In cockroaches there are two slow (177D and 177E) and two fast (178 and 179) coxal depressor muscles, the former being innervated by the slow motoneuron D_s and the latter by D_f; some fibers in 177D also receive inputs from D_f [41]. With this and the discussion of section 3.2 in mind, we now develop our network model.

4. A hexapedal neuro-motor complex.

Pearson did not address contralateral connectivity, but it is natural to extend his model to a network of six mutually inhibiting units, as shown in Figure 2(c) (also cf. the stick-insect pattern generator proposed in [14, Fig. 4]). This architecture promotes contra- and ipsilateral neighbors to burst in antiphase, leading units $1, 2, 3$ and $4, 5, 6$ to form two groups, internally in phase but mutually in antiphase, thus forming the left and right (depressor) tripods. As noted above, we do not include interneurons, so the output of each CPG neuron inhibits the slow and fast depressor motoneurons directly. By inhibiting motoneuronal activity, the CPG selects both a stepping pattern and sets the leg cycle frequency, but the CPG spiking frequency does not directly affect motoneuron spiking frequencies, which are jointly adjusted by the local proprioceptive feedback and CNS drive; see Figures 2(a), (b). CPG neurons and both slow and fast motoneurons will be modeled by (2.1)–(2.2) with differing parameters as specified in Table 1.

Inhibitory coupling can be achieved via synapses that produce negative postsynaptic currents, or presynaptically by depressing a synapse. Lacking more precise information, we choose the former mechanism. Following [47, p. 15], [48, p. 180], we adopt the first order dynamics

$$(4.1)\qquad \dot{s} = \alpha\, G(v_{\mathrm{pre}})\,(1-s) - \beta s, \quad \text{with} \quad G(v_{\mathrm{pre}}) = \frac{T_{\max}}{1+e^{-k_{\mathrm{pre}}(v_{\mathrm{pre}}-E_{\mathrm{syn}}^{\mathrm{pre}})}},$$

in which v denotes the potential of the presynaptic neuron and α, β and the parameters $T_{\max}, k_{\mathrm{pre}}$, and $E_{\mathrm{syn}}^{\mathrm{post}}$ defining the concentration of transmitter release $G(v_{\mathrm{pre}})$ set the timescale of the synaptic rise and decay described by the nondimensional variable s. The variables s enters the postsynaptic cell in the first equation of (2.1) as an additional term,

$$(4.2)\qquad C\dot{v} = -[I_{\mathrm{Ca}} + I_{\mathrm{K}} + I_{\mathrm{L}} + I_{\mathrm{KS}}] + I_{\mathrm{ext}} - \bar{g}_{\mathrm{syn}}\, s \cdot (v - E_{\mathrm{syn}}^{\mathrm{post}}),$$

where $\bar{g}_{\mathrm{syn}}$ denotes synaptic strength and the current $I_{\mathrm{syn}} = -\bar{g}_{\mathrm{syn}}\, s \cdot (v - E_{\mathrm{syn}}^{\mathrm{post}})$ induced in the postsynaptic cell is typically positive and hence depolarizing (resp., negative and hence hyperpolarizing) for excitatory (resp., inhibitory) synapses [33]. A different form of the s-equation (4.1) appears in [49]. We have checked that this produces similar results to those described below.

Table 1 lists parameter values adopted for the CPG and motoneuron models and Table 2 lists those for the synapses (standard inhibitory GABA_A; see [47]). All three types of neurons have equal "fixed" parameter values except for C, ϵ, and E_{syn}. The physiologically adjustable control parameters, $\bar{g}_{\mathrm{KS}}$, I_{ext}, and $\bar{g}_{\mathrm{syn}}$ with nominal "standard" values indicated by asterisks will be varied to match the data summarized in section 3. To obtain equal current injection into all six CPG neurons under stationary conditions, we chose half weights for the synapses from units 1 and 3 to 5, and 4 and 6 to 2 (i.e., $\bar{g}_{\mathrm{syn}} = 0.005$ in place of 0.01 as given in Table 2), since the middle leg units receive input from three others, while other units have inputs from only two (see Figure 2(c)). Contralateral CPG synapses are set at full strength, since in-phase contralateral activity can occur with weak inhibition; cf. [5] and see sections 5.3 and 5.7 below.

Table 1

Parameters for CPG and fast and slow motoneurons D_f, D_s. Maximal conductances are expressed in mS/cm^2, the reversal and threshold potentials in mV, the slope coefficients in mV/s, and the capacitance C in $\mu F/cm^2$.

	C	$\bar{g}_{\mathrm{Ca}}$	$\bar{g}_{\mathrm{K}}$	$\bar{g}_{\mathrm{KS}}$	$\bar{g}_{\mathrm{L}}$	E_{Ca}	E_{K}	E_{L}	$v_{th_{\mathrm{Ca}}}$	$v_{th_{\mathrm{K}}}$	v_{th_c}
CPG	1.2	4.4	9.0	0.19*	2.0	120	−80	−60	−1.2	2	−27
D_f	1.39	4.4	9.0	0.25*	2.0	120	−80	−60	−1.2	2	−27
D_s	2.4	4.4	9.0	0.50	2.0	120	−80	−60	−1.2	2	−27

	$k_{0_{\mathrm{Ca}}}$	$k_{0_{\mathrm{K}}}$	k_c	ϵ	δ	I_{ext}
CPG	0.056	0.1	0.8	4.9	0.052	35.6*
D_f	0.056	0.1	0.8	4.18	0.044	36.3*
D_s	0.056	0.1	0.8	2.0	0.0002	50*

Table 2

Synapse parameters. Only the CPG neurons have "outgoing" synapses.

	$E_{\mathrm{syn}}^{\mathrm{pre}}$	$E_{\mathrm{syn}}^{\mathrm{post}}$	k_{syn}	α	β	$\bar{g}_{\mathrm{syn}}$	$T_{\max}$
CPG	2	−70	0.22	5000	0.180	0.01*	$2 \cdot 10^{-3}$
D_f		−70	0.22	5000	0.180	0.2*	
D_s		−70	0.22	5000	0.180	0.9*	

4.1. Pairs of coupled bursting neurons.

Before studying the full circuit of Figure 2(c), we consider a pair of CPG neurons with mutually inhibitory and excitatory couplings and a CPG neuron unidirectionally coupled to fast and slow motoneurons.

Depending on their intrinsic bursting frequencies and the strength of the coupling term $\bar{g}_{\mathrm{syn}}$ of (4.2), the units may entrain (frequency lock). Figures 3(a), (b) show pairs of identical CPG neurons mutually coupled by inhibitory synapses (left column) and excitatory synapses (right column). In the first case they antiphase lock within a cycle; in the second the bursts entrain, although individual spikes may not. Unidirectionally driven fast motoneurons entrain to the bursting frequency of CPG neurons in Figures 3(g), (h). Slow motoneurons are essentially continual spikers, but with sufficiently strong inhibitory coupling, they can be made to burst in alternation with the CPG inputs in agreement with animal recordings; see Figure 3(i). With excitatory coupling, spiking persists throughout, but the rate increases during an incoming CPG burst; see Figure 3(j). The intervening panels (c)–(f) show the synaptic variable s and the resulting currents $-I_{\mathrm{syn}}$. The s-dynamics is similar in both inhibitory and excitatory cases; the major difference lies in postsynaptic currents.

4.2. A hexapedal CPG.

We now move to the full circuit of Figure 2(c). Synaptic currents and other relevant parameters will be distinguished for CPG, fast, and slow motoneurons by adding appropriate subscripts. Each CPG neuron forms three types of synapses: to other CPG neurons through $I_{\mathrm{syn,CPG}}$ and to fast and slow motoneurons through $I_{\mathrm{syn,D}_f}$, $I_{\mathrm{syn,D}_s}$, respectively. Figure 4 shows typical time histories of ipsilateral and contralateral CPG neurons and motoneurons; note the alternating activity of the left $(1, 2, 3)$ and right $(4, 5, 6)$ tripods. Also, burst durations of the slow motoneurons D_s are longer than those of CPG neurons, and

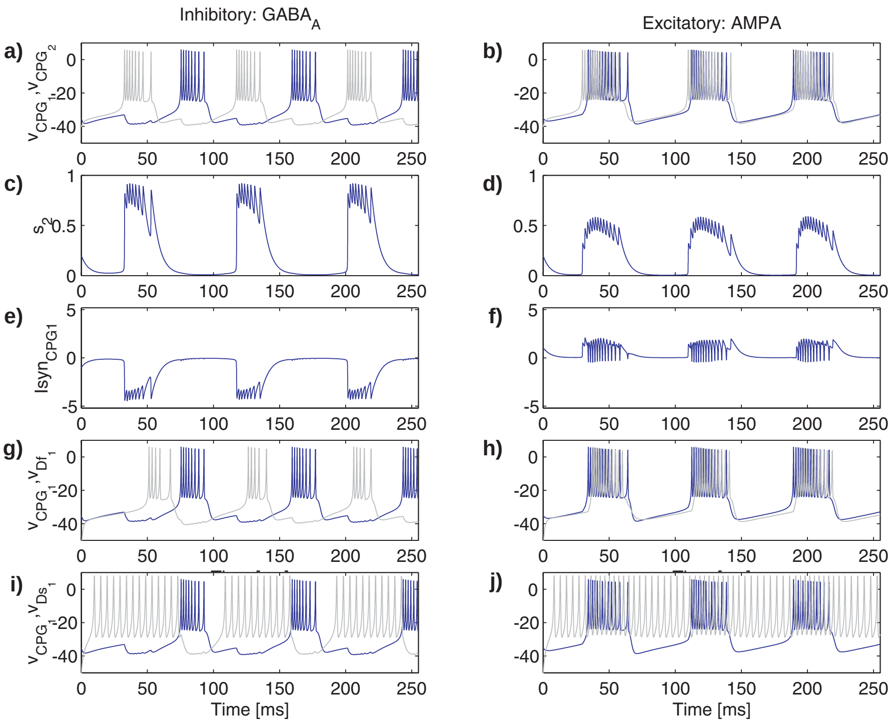


Figure 3. *Two coupled bursting neurons, with GABA$_A$ inhibitory synapses (left panels) and AMPA excitatory synapses (right panels). Panels* (a)–(b) *show membrane voltages of mutually coupled CPG$_1$ (dark grey) and CPG$_2$ (light grey) neurons, respectively; inhibitory coupling causes antiphase bursts* (a) *and excitatory coupling causes in-phase bursts* (b)*. Panels* (c)–(d) *show synaptic dynamics $s(t)$ and* (e)–(f) *the resulting synaptic currents $-I_{syn}$: Negative in the inhibitory case and positive in the excitatory case. Panels* (g), (h) *show the membrane voltage of a CPG neuron (dark grey) superimposed on that of a unidirectionally coupled fast motoneuron D_{f_1} (light grey). With inhibitory coupling* (g), *D_{f_1} bursts in antiphase with respect to CPG, but in-phase with excitatory coupling* (h)*. Panels* (i)–(j) *show the membrane voltage of a CPG neuron (dark grey) superimposed on that of a unidirectionally coupled slow motoneuron D_{s_1} (light grey). With inhibitory coupling* (i), *D_{s_1} bursts in antiphase with respect to CPG, but with excitatory coupling, it continues to spike, with increased rate during CPG bursts* (j)*. Parameters are as in Table* 1 *except for $I_{CPG} = 36.3$, with coupling strengths $\bar{g}_{CPG,CPG} = 0.15$, $\bar{g}_{CPG,D_f} = 0.25$, and $\bar{g}_{CPG,D_s} = 0.4$ for both the inhibitory and excitatory cases. The coefficients for AMPA synapses are as in Table* 2*, except for $\alpha = 1100$, $\beta = 0.190$, $E_{syn}^{post} = 0$. Some panels show effects of transients, and we note that while bursts are synchronized, individual spikes (and spike numbers) need not be.*

their spiking frequency is approximately constant, in agreement with experiments [26].

Via $I_{\text{ext}} = I_\alpha$ and $\bar{g}_{\text{KS},\alpha}$, where $\alpha = \{\text{CPG}, \text{D}_f, \text{D}_s\}$, we can adjust the stepping frequency, the number of APs in D$_f$, the spiking rate of D$_s$, and their duty cycles, as described in [1] and section 2 above. We note that all three neuron types can have different duty cycles. In the locomotion literature duty cycle normally refers to S-type (slow) muscle fibers or slow motoneuron activity (in fast fibers it is not a relevant measure). In our network, the duty cycle of the slow motoneurons can be indirectly controlled through that of the CPG neurons. Since CPG neurons drive motoneurons through inhibitory synapses, and duty cycles of the former are generally less than 0.5, motoneuron duty cycles typically exceed 0.5. Hence, by suitable

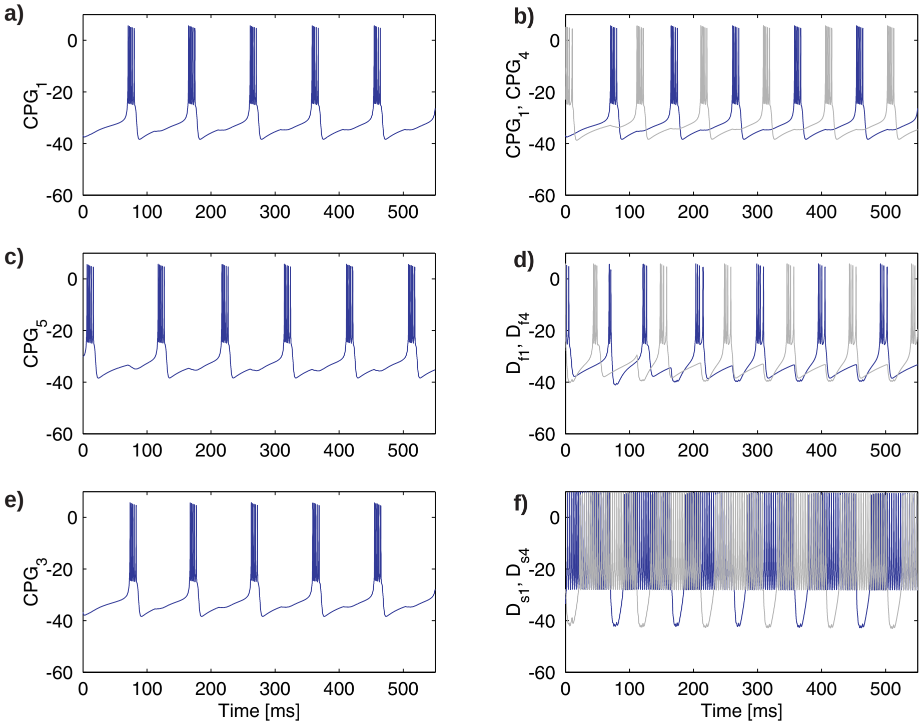


Figure 4. *Membrane voltages of CPG neurons and motoneurons in the hexapedal model during fictive locomotion. Neurons are labeled as in Figure* 2(c)*, and parameter values are as in Tables* 1 *and* 2*. The left column shows ipsilateral CPG neurons, while the right column shows contralateral CPG neurons and fast and slow motoneurons for units* 1 *and* 4*. Bursting frequency is* 10.4 *Hz, D_f has* 4 *APs per burst, and D_s spikes at a rate of* ≈ 290 *Hz with a duty cycle of* 0.80*;* $(f_{Burst}, n_{AP}, f_{Spike}, \Delta_{D_s}) = (10.4$ *Hz*$, 4, 287$ *Hz*$, 0.80)$*. Parameters are as in Tables* 1 *and* 2*.*

parameter choices we can reproduce Pearson's finding that motoneuron burst durations vary from 0.4 to 0.9 of the full cycle period as the latter increases; see Figure 6(d). From now on, unless otherwise stated, by duty cycle we mean the D_s duty cycle.

In the following we show how the network can be adjusted for different locomotive requirements, in comparison with the nominal case of Figures 4(b), (d), (f) in the insect's preferred speed range, in which the stepping frequency is 10.3 Hz, the D_f have four APs per burst, and the D_s spike at a rate of 279 Hz and have a duty cycle of 0.59. We write these four "outputs" as $(f_{\mathrm{Burst}}, n_{\mathrm{AP}}, f_{\mathrm{Spike}}, \Delta_{\mathrm{D}_s})$. Figures 5($\mathrm{a}_1$)–($\mathrm{a}_3$) show slow walking (3.66 Hz, 1, 147 Hz, 0.88). Figures 5(b_1)–(b_3) show how it is possible to vary the number of APs in fast motoneurons and the spiking frequency of slow motoneurons independent of stepping frequency: still slow walking but with increased force production, as required, e.g., for hill climbing (3.66 Hz, 7, 353 Hz, 0.95). Figures 5(c_1)–(c_3) show fast stepping, with an intermediate number of APs and spiking rate (17.2 Hz, 4, 287 Hz, 0.44). Note that in this case the D_s duty cycle is slightly less than 0.5 and there is no overlap.

To adjust to rapid external disturbances, flexibility is required in load as well as speed. In the following we show how, in a multiparametric setting, the four main characteristics of

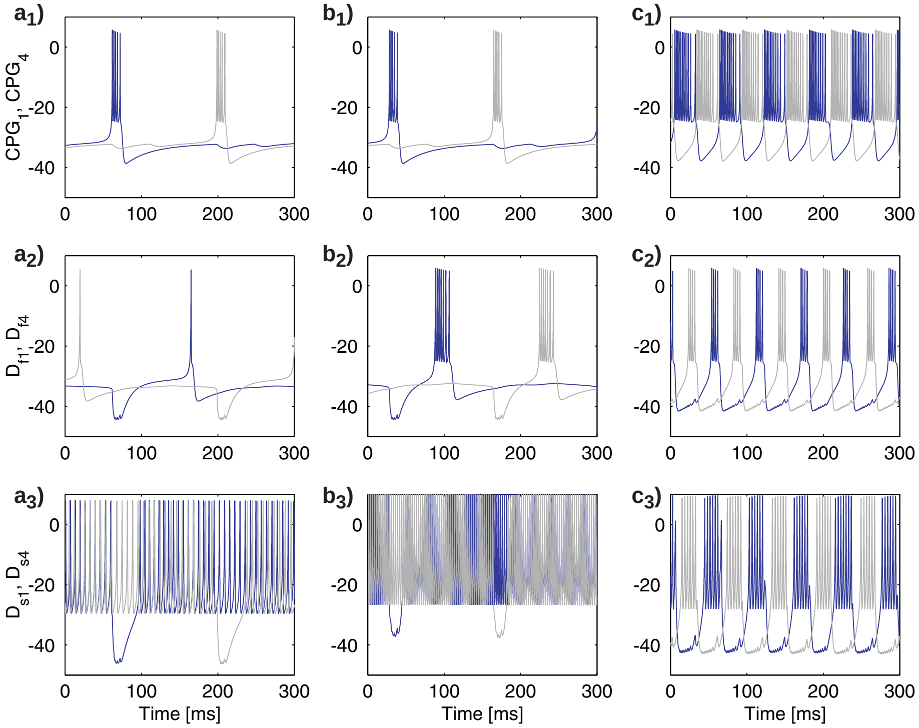


Figure 5. *Membrane voltages of CPG neurons and fast and slow motoneurons in the hexapedal model during fictive locomotion, for comparison with "nominal" case of Figure 4 (right column).* (a_1)–(a_3) *Slow stepping, low force:* $(f_{Burst}, n_{AP}, f_{Spike}, \Delta_{D_s}) = (3.66\ Hz,\ 1,\ 147\ Hz,\ 0.88)$*; parameters are as in Figure 4 except for* $I_{CPG} = 35.38$*,* $I_{D_f} = 35.7$*,* $I_{D_s} = 44$*,* $\bar{g}_{KS,D_f} = 0.45$*,* $\bar{g}_{CPG,D_f} = 0.5$*.* (b_1)–(b_3) *Slow stepping, large force:* $(f_{Burst}, n_{AP}, f_{Spike}, \Delta_{D_s}) = (3.66\ Hz,\ 7,\ 353\ Hz,\ 0.95)$*; parameters are as in Figure 4 except for* $I_{CPG} = 35.38$*,* $I_{D_f} = 35.8$*,* $I_{D_s} = 64$*,* $\bar{g}_{KS,D_f} = 0.13$*,* $\bar{g}_{CPG,D_f} = 0.5$*.* (c_1)–(c_3) *Fast stepping, medium force:* $(f_{Burst}, n_{AP}, f_{Spike}, \Delta_{D_s}) = (17.2\ Hz,\ 4,\ 287\ Hz,\ 0.44)$*. Parameters are as in Figure 4 except for* $I_{CPG} = 38.4$*,* $I_{D_f} = 37.4$*,* $I_{D_s} = 50$*.*

the network can be adjusted over a wide range, with sufficient independence. Parameters not explicitly noted are as in Tables 1–2. Figure 6(a) shows the variation of bursting frequency with I_{CPG} parametrized by the maximal conductance $\bar{g}_{\mathrm{KS,CPG}}$. Together they span the range 5–26 Hz (although a lower frequency of 3.2 Hz was obtained with $I_{\mathrm{CPG}} = 35.38$, $I_{\mathrm{D}_f} = 35.7$), encompassing the entire range over which *Blaberus discoidalis* uses the double-tripod gait. Figure 6(c) shows how the duty cycle of slow motoneurons is affected by changes in I_{CPG} and $\bar{g}_{\mathrm{KS}}$. Figure 6(b) shows the variation of the spiking frequency of the slow motoneurons with I_{D_s} parametrized by I_{CPG}. Variation of I_{D_s} in the range 38–64 provides frequencies from 124 Hz to 389 Hz. Figure 6(c) shows duty cycle variation with I_{CPG} parametrized by $\bar{g}_{\mathrm{KS,CPG}}$, indicating coverage of the range from 0.4 to 0.9, and Figure 6(d) shows this data superimposed on measurements of Pearson [26].

Figure 6(e) shows the variation of the number of APs in fast motoneurons with I_{D_f} parametrized by I_{CPG}. In the first case the number of APs per burst changes only from 3 to 4, but this is significant in the 10 Hz frequency range; cf. [37]. A wider range is obtained when $I_{\mathrm{CPG}} = 38.4$, corresponding to a stepping frequency of 17.0 Hz; here the number of APs

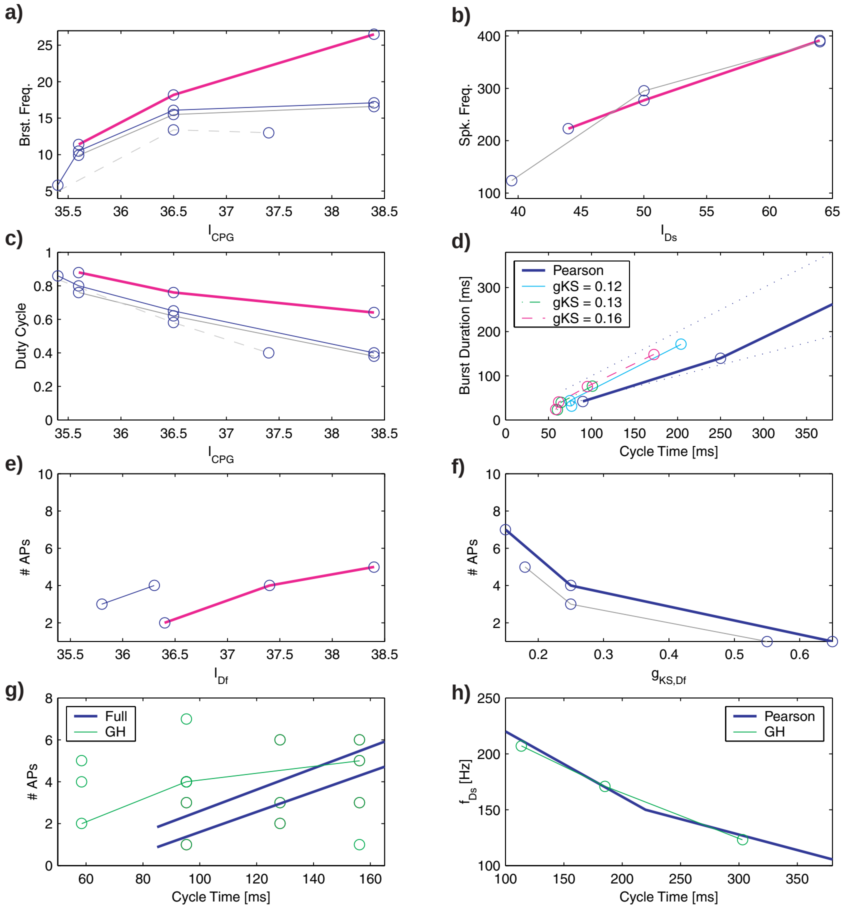


Figure 6. (a) *Network bursting frequency vs.* I_{CPG} *for* $\bar{g}_{KS,CPG} = 0.35$ *(bold),* 0.19 *(dark),* 0.18 *(solid grey), and* 0.15 *(dashed grey).* (b) *Slow motoneuron spiking frequency vs.* I_{D_s} *for low* $I_{CPG} = 35.6$ *(solid, stepping frequency* 10.5 *Hz), and* 38.4 *(bold, stepping frequency* 17.0 *Hz).* (c) *Duty cycle vs.* I_{CPG}*, parametrized by* $\bar{g}_{KS,CPG}$*, parameter values and curves as in panel* (a). (d) *Duty cycle vs. cycle time: Data from Pearson* [26, *Fig.* 6] *(bold) and obtained by varying* I_{CPG} *from* 35.38 *to* 36.7 *and keeping* $\bar{g}_{KS}$ *fixed at the values indicated (dashed and broken lines); dotted lines correspond to* 50% *and* 100% *duty cycle.* (e) *Number of fast motoneuron APs per burst vs.* I_{D_f}*, for* $I_{CPG} = 35.6$ *(solid, bursting frequency* 10.5 *Hz);* $I_{CPG} = 38.4$ *(bold, bursting frequency* 17.0 *Hz).* (f) *Number of fast motoneuron APs per burst vs.* g_{KS,D_f} *for* $I_{CPG} = 35.41$*,* $I_{D_f} = 35.7$ *(solid, bursting frequency* 6.4 *Hz);* $I_{CPG} = 35.6$*,* $I_{D_f} = 36.5$ *(bold, bursting frequency* 10.5 *Hz).* (g) *Fast motoneuron APs vs. cycle time: Data from Full* [37, *Fig.* 5] *with mesothoracic (upper, bold) and metathoracic (lower, bold) regression lines. Model results shown in circles and broken line; see text for explanation.* (h) *Slow motoneuron spiking frequency vs. cycle time: Data from Pearson* [26, *Fig.* 7] *(bold) and model results on broken line; see text.*

ranges from 2 to 5. Figure 6(f) shows variation of the number of APs with $\bar{g}_{\mathrm{KS,D}_f}$ parametrized by I_{CPG}, indicating that from 1 to 7 APs can be delivered. Recalling that each spike causes a muscle fiber twitch, the model can therefore achieve up to a sevenfold graded increase in force production, covering the entire range described in [37]. Finally, Figures 6(g), (h) replot the fast motoneuron AP numbers and slow motoneuron spike rates achieved by the model in comparison with those measured by Full et al. [37] and Pearson [26], showing that the model can reproduce the data rather well. To match Full's overall finding that slower stepping (increased cycle time) results in more APs, we adjusted the network bursting frequency via $I_{\text{CPG}} = 35.41$, 35.6, 38.4 and concurrently adjusted the bias current $I_{\text{D}_f} = 35.7$, 36.3, 36.4 to produce the broken line (for the rightmost data point $g_{\text{KS,D}_f} = 0.18$ was slightly less than for the others $g_{\text{KS,D}_f} = 0.19$). The circles in panel (g) show network behaviors obtained for a broader variation of I_{CPG}, I_{D_f}, and $\bar{g}_{\text{KS,D}_f}$ that span the wide variability identified in [37, Fig. 5]. The circles in (h) were obtained by concurrently changing I_{CPG} from 35.38 to 35.6 and I_{D_s} from 40 to 44. We also found other cases in which the number of APs decreases with increasing cycle time.

All results shown in this section correspond to tripod gaits, with 1:1 entrainment of CPG and fast and slow motoneurons, as in Figure 5. The impossibility of extending some curves beyond the ranges shown (e.g., Figure 6(d), solid line) is due to failure of "normal" network properties: e.g., CPG neurons, fast and/or slow motoneurons cease to fire at all (typically at low values of the current), or fire tonically (high bias currents); or 1:1 phase locking of fast motoneurons and CPG neurons is lost. Nonetheless, these simulations show that the tripod gait can be maintained over a wide range of speeds and duty cycles, and that in a multiparameter setting, bursting frequency, spiking frequency, duty cycle, and the number of APs can be almost independently changed.

5. Reduction to phase oscillators.

In this section we review the phase reduction and averaging methods and apply them to coupled bursting CPG neurons of the type (2.1) with synaptic dynamics (4.1). We derive reduced sets of ODEs describing mutually coupled pairs of neurons and the CPG network of Figure 2(c) in terms of relative phases, and analyze them to find phase locked solutions and their stability properties.

5.1. The phase response curve.

We write the ODE for a single cell in the compact form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \alpha\mathbf{p}(\mathbf{x}, t, \dots); \quad \mathbf{x} \in \mathbb{R}^n, \tag{5.1}$$

where $\mathbf{p}$ denotes the coupling function (of strength α) and $(\dots)$ in the argument of $\mathbf{p}$ contains state variables of all cells that synapse onto the one in question, as well as external inputs. The phase reduction method originated in work of Malkin and Winfree [50, 51]; more details can be found in [52, Chap. 9] and [53], and an application to the "standard" Hodgkin–Huxley equations in [54].

We assume that, for $\alpha = 0$, (5.1) has an attracting hyperbolic limit cycle Γ_0, with period T_0 and frequency $\omega_0 = \frac{2\pi}{T_0}$ (the bursting cycle). We define a scalar phase variable $\phi(\mathbf{x}) \in [0, 2\pi)$ for all $\mathbf{x}$ in some neighborhood U of Γ_0 (within its domain of attraction), such that the phase evolution has the simple form $\dot{\phi} = \omega_0$ for all $\mathbf{x} \in U$. The cycle Γ_0 persists for small $\alpha \neq 0$ [21], and, from the chain rule, we deduce that

$$\dot{\phi} = \frac{d\phi}{dt} = \frac{\partial\phi}{\partial\mathbf{x}} \cdot [\mathbf{f}(\mathbf{x}) + \alpha\mathbf{p}(\mathbf{x}, t, \dots)] = \omega_0 + \alpha\frac{\partial\phi}{\partial\mathbf{x}} \cdot \mathbf{p}(\mathbf{x}, t, \dots). \tag{5.2}$$

Equation (5.2) defines a first order PDE that the scalar field $\phi(\mathbf{x})$ and its inverse $\mathbf{x} = \mathbf{x}(\phi)$ must satisfy; $\phi(\mathbf{x})$ is unique up to a translational constant which may be fixed by setting $\phi(\mathbf{x}) = 0$ at a distinguished point of Γ_0. For periodically spiking neurons, this is often the voltage peak; in the present application it will be the upward crossing of -30 mV preceding the first spike in the burst. The theory of isochrons [55] implies that the phase space $\mathbb{R}^n$ near Γ_0 is foliated by $(n-1)$-dimensional manifolds

$$M_{\bar{\phi}} = \Big\{ \mathbf{x} \in \mathcal{B} : \lim_{t\to\infty} \mathbf{x}(t) \sim \phi(t) = \omega_0 t + \bar{\phi} \Big\},$$

from which solutions approach Γ_0 with the same asymptotic phase.

Introducing the relative phase $\psi = \phi - \omega_0 t$ and approximating the derivative in (5.2) by its value on the uncoupled limit cycle $\mathbf{Z}(\phi) \overset{\text{def}}{=} \frac{\partial \phi}{\partial \mathbf{x}}|_{\Gamma_0(\phi)}$, (5.2) becomes

$$\dot{\psi} = \alpha \mathbf{Z}(\phi) \cdot \mathbf{p}(\phi). \tag{5.3}$$

For mutual coupling among N identical units, defining the phase variables $\mathbf{x}_i = \mathbf{x}_i(\phi_i)$ and $\psi_i = \phi_i - \omega_0 t$, this generalizes to

$$\dot{\psi}_i = \sum_{j\neq i}^{N} \alpha_{ji} \mathbf{Z}(\phi_i) \cdot \mathbf{p}_{ji}(\phi_i, \phi_j). \tag{5.4}$$

For weak coupling ($|\alpha| \ll 1$), the phases ϕ_i evolve on a much faster time scale than ψ_i, so we may appeal to averaging theory [21] (cf. [52, p. 259, Malkin's theorem]) to integrate over the unperturbed period and obtain

$$\dot{\psi}_i = \sum_{j\neq i}^{N} \alpha_{ji} H_{ji}(\psi_i - \psi_j), \tag{5.5}$$

where

$$H_{ji}(\psi_i - \psi_j) = \frac{1}{T_0} \int_0^{T_0} \mathbf{Z}(\omega_0 t + \psi_i) \cdot \mathbf{p}_{ji}(\Gamma_0(\omega_0 t + \psi_i), \Gamma_0(\omega_0 t + \psi_j)) dt. \tag{5.6}$$

Note that only phase *differences* appear in the averaged coupling functions H_{ji} due to periodicity of the integrand in (5.6).

In the case that the perturbation or coupling functions $\mathbf{p}_{ji}$ only enter through the first component of $\mathbf{x}$, as in $\dot{v}$ via I_{syn} in (4.2), we have

$$\alpha_{ji} \mathbf{p}_{ji}(\mathbf{x}_i, \mathbf{x}_j, t) = \begin{bmatrix} \bar{g}_{\text{syn},ji}\, s_j \cdot (v_i - E_i) \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{5.7}$$

where s_j denotes the synaptic variable associated with the jth cell and $\bar{g}_{\text{syn},ji}$ the synaptic strength from the jth to the ith cell. Here only the first component $Z_1(\phi)$ of the 2π-periodic function $\mathbf{Z}(\phi)$ appears in (5.6); this is called the *phase response curve* (PRC), and it may be approximated numerically by perturbing from the limit cycle at each phase ϕ with a voltage increment $v \mapsto v + \Delta v$ and allowing the solution to recover to its new asymptotic phase $\phi \mapsto \phi + Z_1(\phi)$. The resulting *infinitesimal PRC* is valid in the combined limit

$$Z_1(\phi) = \lim_{\substack{\Delta v \to 0 \\ t \to \infty}} \frac{\Delta \phi}{\Delta v}. \tag{5.8}$$

$Z_1(\phi)$ may also be computed by use of adjoint theory [52], e.g., as implemented in the software XPP [56]. For $Z_1(\phi) > 0$ (resp., < 0), positive voltage perturbations advance (resp., retard) the phase.

5.2. PRCs and averaged coupling functions. Applying the theory sketched above to (2.1), we obtain the infinitesimal PRC of Figure 7(a). Here $Z_1(\phi)$ was computed numerically by taking successively smaller voltage perturbations Δv starting at $\Delta v = 31.6$ mV and reducing to 0.1 mV; the PRC stabilized in the form shown for $|\Delta v| < 1$ mV. If the linearity assumption inherent in (5.8) holds, positive and negative perturbations should yield the same result, since the infinitesimal PRC is a derivative. We verified that this is the case using perturbations $\Delta v = \pm 0.08$ and finding good convergence over the whole range $\theta \in [0, 2\pi)$; see the solid and dashed curves in Figure 7(a).

In Figure 7(b) we show how the estimate of Z at $\theta = 30.6^{\text{o}}$ changes as perturbation size increases. This indicates how weak the coupling should be for the theory to hold, i.e., the maximum size of α allowed in (5.2)–(5.3). As $|\Delta v|$ increases, linearity is lost in three different respects: (i) (for a given θ) the phase difference $\Delta\phi$ is no longer proportional to $|\Delta v|$; (ii) positive and negative perturbations give different contributions; (iii) strong nonlinear effects appear in the perturbed limit cycle; spikes can be deleted, as in Figure 7(c) for $\Delta v = -7.5$ mV, and spikes or entire bursts can be added, as in Figure 7(d) for $\Delta v = +7.5$ mV. The threshold for "small" perturbations may depend on θ and on specific parameters (see comment in section 5.4). In Figure 7(e) we show the estimate of Z at $\theta = 168.8^{\text{o}}$; for negative perturbations, linearity is maintained up to $h = -31.6$ mV but is lost at around $h = +7$ mV for positive perturbations. This could imply that antiphase solutions are more robust than in-phase solutions; see the discussion in section 5.4.

We observe three distinct regions in the PRC. During the burst, sensitivity to each spike is evident, with maximal sensitivity to the final one. After this, there is a period of relative insensitivity, followed by a region dominated by a large smooth phase advance. This third region is largely unaffected by changes in bursting frequency, duty cycle, or number of APs in the burst, since the end of the cycle remains very similar; see also Figures 10(b_1)–(b_2) and 11(a_1)–(a_2), below. A PRC of similar form was derived experimentally by Delcomyn [57] for the locust flight CPG.

The inability of the infinitesimal PRC to represent the loss or addition of spikes and bursts follows from the tight coiling of the limit cycle of (2.1) in phase space. This implies that the isochronic manifolds $M_{\bar{\phi}}$ are globally convoluted, and (moderate) perturbations exist that can skip or repeat spikes by taking the perturbed voltage $v + \Delta v$ from one spike to a point near another. Hence phase reduction must be used with care, although if the number of spikes within a burst is large compared to the burst period, skipping or adding spikes will not greatly affect the averaged coupling functions H_{ji} of (5.5)–(5.6). In what follows synaptic conductances are small enough to remain within the range of infinitesimal PRC validity.

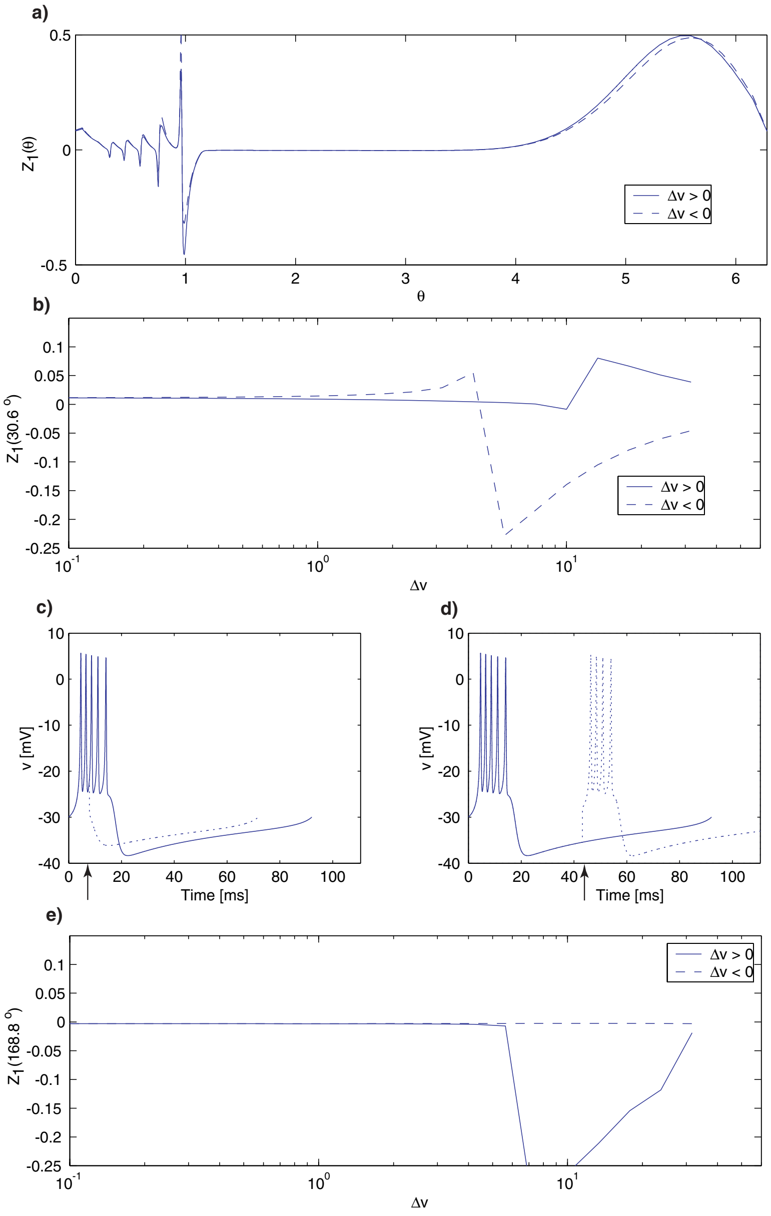


Figure 7. *The infinitesimal PRC of* (2.1) *for the standard parameter set of Tables* 1–2. (a) $Z_1(\phi)$ *computed with a perturbation of* $\Delta v = \pm 0.08$ *mV (solid and dashed, respectively).* (b) $Z_1(30.6^o)$ *as a function of perturbation size and sign:* $\Delta v > 0$ *(solid),* $\Delta v < 0$ *(dashed).* (c) *The unperturbed (solid) and perturbed (dashed) cycles with* $\Delta v = -7.5$ *mV applied at* $\theta = 30.6^o$ *(arrow): Spikes can be removed.* (d) *The unperturbed (solid) and perturbed (dashed) cycles for* $\Delta v = +7.5$ *mV applied at* $\theta = 168.8^o$ *(arrow): Spikes or entire bursts, as here, can be added.* (e) $Z_1(168.8^o)$ *as a function of perturbation size and sign:* $\Delta v > 0$ *(solid),* $\Delta v < 0$ *(dashed).*

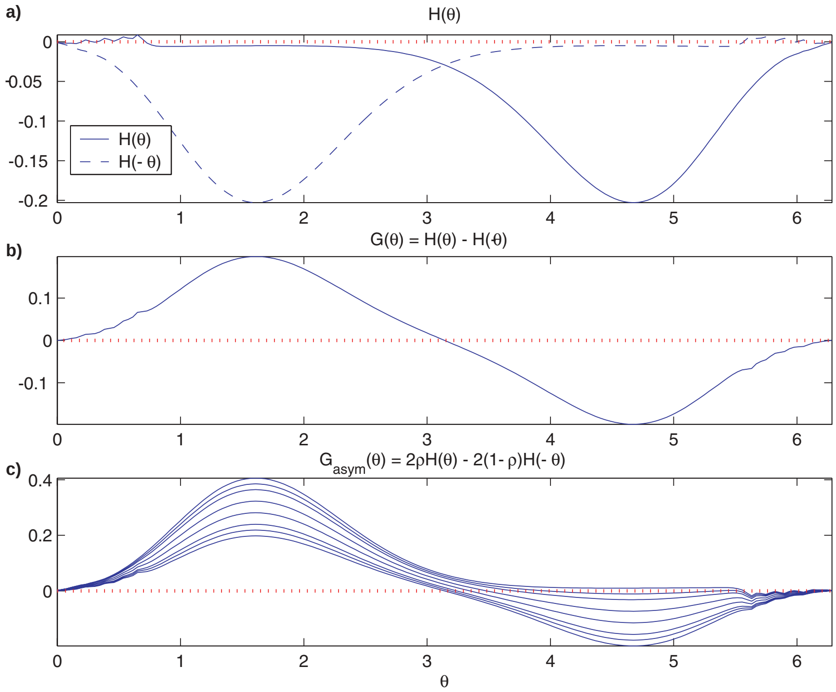


Figure 8. (a) *The averaged coupling function $H(\theta)$ (solid) for an inhibitory synapse; $H(-\theta)$ also shown (dash-dotted).* (b) *The phase difference coupling function $G(\theta) = H(\theta) - H(-\theta)$. Note that $G'(0) > 0 > G'(\pi)$.* (c) *Asymmetric coupling: $G(\theta) = 2\rho H(\theta) - 2(1-\rho)H(-\theta)$; ρ varies from 0 to 0.5 from top to bottom curve.*

5.3. Coupled bursting neurons.

We now return to study coupled bursters in a phase-reduced setting. From Figures 2(c), (d), two types of coupling appear: motoneurons are unilaterally driven by the CPG neurons, whereas CPG neurons are bi- or trilaterally coupled.

Fast motoneurons, unilateral coupling. Allowing for different intrinsic frequencies $\omega_0 + \epsilon_j$, the PRC and averaging theory of section 5.1 lead to phase-reduced dynamics for a CPG neuron ψ_1 and a fast motoneuron ψ_2 of the form

$$\dot{\psi}_1 = \epsilon_1, \quad \dot{\psi}_2 = \epsilon_2 + \alpha H(\psi_2 - \psi_1). \tag{5.9}$$

Phase locking occurs when $\dot{\theta} = \dot{\psi}_1 - \dot{\psi}_2 = 0$:

$$\epsilon_1 - \epsilon_2 = \alpha H(\psi_2 - \psi_1). \tag{5.10}$$

As shown in Figure 8(a) H is almost always negative for the nominal case. Phase reduction therefore predicts that locking can only occur if $\epsilon_1 < \epsilon_2$, i.e., when the intrinsic CPG bursting frequency is lower than that of the motoneurons. This is illustrated in Figures 14(a), (b) below, where $f_{\text{CPG}} = 10.5$ Hz $< f_{\text{D}_f} = 15.4$ Hz. On the contrary, phase locking does not occur when $f_{\text{CPG}} = 18.1$ Hz $> f_{\text{D}_f} = 15.4$ Hz; see Figures 14(c), (d). Indeed, inhibitory coupling delays the bursts; therefore, it is natural to expect that for unilateral coupling the driving neuron should be slower than the follower. If that were not the case, the already slower "follower" would be further slowed, preventing 1:1 phase locking. In fact, for the simulations reported here, 2:1 locking occurs; see Figure 14(d). This could explain the phenomenon of "double bursting" described in [58].

Thus, provided their bursting frequencies are chosen appropriately, fast motoneurons follow CPG neurons and so need not be explicitly included in the reduced analysis of locomotion rhythms to follow. Similarly, slow motoneurons need not be included in a reduced phase description, since they also phase lock via synaptic depression and their duty cycles are determined by those of the CPG neurons (cf. Figures 3(i), 6(c)).

CPG neurons, mutual coupling. For mutual coupling between two identical CPG neurons the reduced phase equations (5.5) become

$$\dot{\psi}_1 = \alpha H(\psi_1 - \psi_2), \quad \dot{\psi}_2 = \alpha H(\psi_2 - \psi_1) \tag{5.11}$$

(since $\alpha_{12}H_{12} = \alpha_{21}H_{21}$, here we may drop the subscripts), and subtracting these we may further reduce to a single scalar ODE for the phase difference $\theta = \psi_1 - \psi_2$:

$$\dot{\theta} = \alpha[H(\theta) - H(-\theta)] \overset{\text{def}}{=} G(\theta). \tag{5.12}$$

Fixed points of (5.12) occur at $H(\theta) = H(-\theta)$, and since H is 2π-periodic, we have $G(\pi) = \alpha[H(\pi) - H(-\pi)] = \alpha[H(\pi) - H(\pi)] = 0$ as well as $G(0) = 0$, implying that, *regardless of the form* of H, (exact) in-phase and antiphase solutions exist; see Figure 8(b). Note that, for $\bar{\theta} = 0$ and π, the equations in (5.11) become $\dot{\psi}_1 = \dot{\psi}_2 = \alpha H(\bar{\theta})$, so that, unless $H(0) = 0$ and/or $H(\pi) = 0$, coupling does change the common frequency $\dot{\phi} = \omega_0 + \dot{\psi}_i$ of the units, even when phase locking occurs.

The stability of fixed points $\bar{\theta}$ of (5.12) is determined by $\frac{\partial G}{\partial \theta}|_{\bar{\theta}} = 2\alpha H'(\bar{\theta})$. As expected, for inhibitory coupling $\alpha H'(0) > 0 > \alpha H'(\pi)$ (Figure 8(b)), so the in-phase solution $\bar{\theta} = 0$ is unstable and the antiphase solution $\bar{\theta} = \pi$ is stable. Stability of the "full" two-phase system (5.11) is determined by the eigenvalues of the 2×2 matrix obtained by linearizing at $\psi_1 - \psi_2 = \bar{\theta}$:

$$\alpha \begin{bmatrix} H'(\bar{\theta}) & -H'(\bar{\theta}) \\ -H'(\bar{\theta}) & H'(\bar{\theta}) \end{bmatrix}; \tag{5.13}$$

these are 0 and $2\alpha H'(\bar{\theta})$ with eigenvectors $(1,1)^{\mathrm{T}}$ and $(1,-1)^{\mathrm{T}}$, respectively. Hence the dynamics is only neutrally stable to perturbations that advance or retard the phases of both units equally, but the antiphase solution is asymptotically stable to perturbations that disrupt the relative phase $\psi_1 - \psi_2$, as indicated by Figure 3.

Finally we consider asymmetric coupling of the type shown in Figure 2(d), which will be discussed further in the context of phase lags. The phase equations for two identical neurons become

$$\begin{aligned} \dot{\psi}_1 &= 2\rho\alpha H(\psi_1 - \psi_2), \\ \dot{\psi}_2 &= 2(1-\rho)\alpha H(\psi_2 - \psi_1), \end{aligned} \tag{5.14}$$

where ρ measures the degree of asymmetry, $\rho = \frac{1}{2}$ being the symmetric case (5.11). The phase difference is then governed by

$$\dot{\theta} = \alpha[2\rho H(\theta) - 2(1-\rho)H(-\theta)] \overset{\text{def}}{=} G_{\text{asym}}(\theta). \tag{5.15}$$

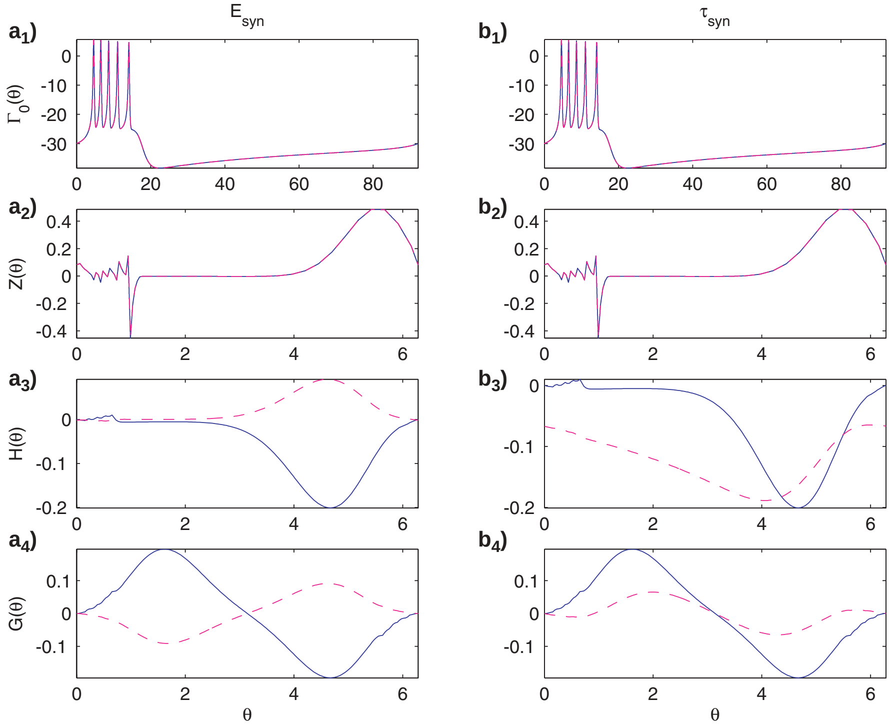


Figure 9. *Dependence of the PRC* $Z(\theta)$*, the averaged coupling function* $H(\theta)$*, and the "difference" function* $G(\theta)$ *on synaptic parameters. Panels* (a_1)–(a_4) *show dependence on synapse type: Inhibitory GABA*$_A$ *(solid) and excitatory AMPA (dashed). Panels* (b_1)–(b_4) *show dependence on synapse timescale: Regular inhibitory GABA*$_A$ *(solid) and slower inhibitory GABA*$_A$ *(dashed) with* $\alpha = 500$, $\beta = 0.018$.

Plots of $G_{\text{asym}}(\theta)$ for different values of $\rho \in [0, 0.5]$ are shown in Figure 8(c). Note that the interior zero occurs at increasing values of $\bar{\theta} > \pi$, representative of a shifted "antiphase" solution. In this case the in-phase solution remains at $\theta = 0$, but this is a special property of the averaged functions H, which vanish at $\theta = 0$.

5.4. Parameter dependence of the PRC and averaged coupling functions.

We will now investigate the effect of certain parameters on the shapes of the functions Z, H, and G. From (5.6), we see that the only parameter that can be factored out is the coupling strength $\alpha_{ji} = \bar{g}_{\text{syn},ji}$, which scales Z, H, and G; the other parameters change their forms more generally.

Changing from inhibitory GABA$_A$ to excitatory AMPA synapses implies changing α, β, and $E_{\text{syn}}^{\text{pre}}$ in (4.1) (cf. caption to Figure 3). As a result, H becomes almost always positive (Figure 9(a$_3$)), but more importantly $G'(\pi)$ becomes positive and $G'(0)$ negative, making the antiphase solution unstable and the in-phase solution stable; see Figure 9(a$_4$). Note that $Z(\theta)$ is unaffected by this change. Even more interesting is the effect of a slower time scale on *inhibitory* synapses. Under a ten-fold increase ($\alpha = 500$, $\beta = 0.018$) the in-phase solution becomes *stable* and two new unstable solutions appear in a pitchfork bifurcation; see Figure 9(b$_4$).

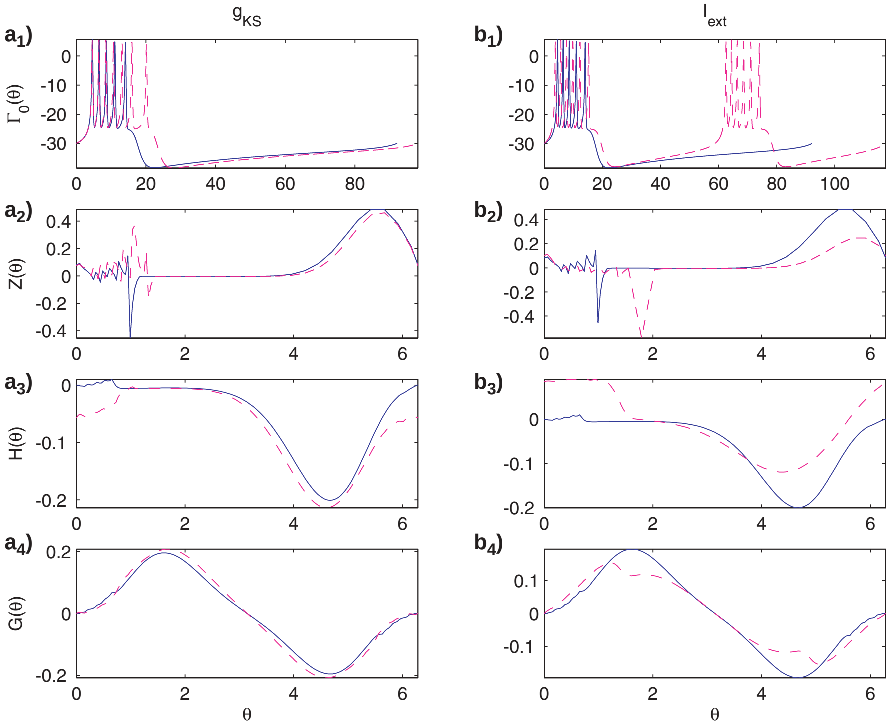


Figure 10. *Dependence of the PRC $Z(\theta)$, the averaged coupling function $H(\theta)$, and the function $G(\theta)$ on $\bar{g}_{KS}$ and I_{ext}. Panels* (a$_1$)–(a$_4$) *show dependence on the maximal conductance $\bar{g}_{KS}$: Nominal case of Tables* 1 *and* 2 *(solid) and decreased value: $\bar{g}_{KS} = 0.16$ (dashed). Panels* (b$_1$)–(b$_4$) *show dependence on the external current: Nominal case (solid) and increased current $I_{ext} = 36.5$ (dashed).*

Changing the maximal conductance $\bar{g}_{\mathrm{KS}}$ can affect the bursting frequency, the duty cycle, and the number of APs per burst. In this case, even though bursting frequency and AP numbers are significantly modified (Figure 10(a$_1$)), and accordingly $Z(\theta)$ and $H(\theta)$ (Figures 10(a$_2$)–(a$_3$)), the net effects on the function $G(\theta)$ largely average out. Results not shown indicate that increasing the maximal conductance to $\bar{g}_{\mathrm{KS}} = 0.25$ can introduce extra spikes, violating the infinitesimal PRC assumption; this case is discussed further as a finite perturbation below. Changing I_{ext} yields analogous results; e.g., in spite of a substantial change in bursting frequency (Figure 10(b$_1$)), the final form of $G(\theta)$ is not greatly modified (Figure 10(b$_4$)).

Finally we show that, even in cases in which spikes are lost or added, the infinitesimal PRC undergoes sharp transitions (Figures 10(b$_1$)–(b$_2$); cf. Figure 7(b)), and the averaged coupling function also changes significantly; both it and the difference function $G(\theta)$ retain similar forms near $\theta = \pi$ (Figures 11(b$_3$), (b$_4$)). Hence we may still deduce stability information regarding antiphase solutions from the reduced description. Indeed, Figure 7(e) shows that $Z_1(\phi \approx \pi)$ is insensitive to the negative perturbation magnitude up to $\Delta v \approx 30$ mV.

5.5. A phase-reduced model of the CPG. Extension of the above reduction of a mutually coupled pair to the network of six CPG neurons in Figure 2(c) is immediate. We again assume identical units, but as noted in section 4 below Table 1, employ different synaptic strengths

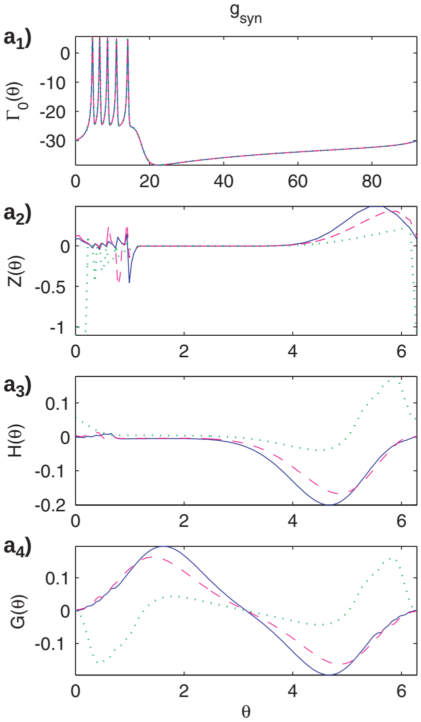


Figure 11. *Dependence of the (infinitesimal) PRC $Z(\theta)$, the coupling function $H(\theta)$, and the function $G(\theta)$ on the size of the perturbation Δv.* (a$_1$)–(a$_4$)*: Nominal case $\Delta v = -0.08$ (solid), medium size $\Delta v = -1$ (dashed), and large $\Delta v = -5$ (dotted). Note that form of Z, H, and G near $\theta = \pi$ remains similar.*

so that all units receive the same net input at steady state. Thus all contralateral connections and ipsilateral connections from units 2 and 5 to 1, 3, 4, and 6 are set at $\bar{g}_{\text{syn}}$, and ipsilateral connections from 1, 3, 4, and 6 to 2 and 5 are set at $\bar{g}_{\text{syn}}/2$. This leads to the following set of six phase equations:

$$\begin{aligned}
\dot{\psi}_1 &= \bar{g}_{\text{syn}} H(\psi_1 - \psi_4) + \bar{g}_{\text{syn}} H(\psi_1 - \psi_5),\\
\dot{\psi}_2 &= \frac{\bar{g}_{\text{syn}}}{2} H(\psi_2 - \psi_4) + \bar{g}_{\text{syn}} H(\psi_2 - \psi_5) + \frac{\bar{g}_{\text{syn}}}{2} H(\psi_2 - \psi_6),\\
\dot{\psi}_3 &= \bar{g}_{\text{syn}} H(\psi_3 - \psi_5) + \bar{g}_{\text{syn}} H(\psi_3 - \psi_6),\\
\dot{\psi}_4 &= \bar{g}_{\text{syn}} H(\psi_4 - \psi_1) + \bar{g}_{\text{syn}} H(\psi_4 - \psi_2),\\
\dot{\psi}_5 &= \frac{\bar{g}_{\text{syn}}}{2} H(\psi_5 - \psi_1) + \bar{g}_{\text{syn}} H(\psi_5 - \psi_2) + \frac{\bar{g}_{\text{syn}}}{2} H(\psi_5 - \psi_3),\\
\dot{\psi}_6 &= \bar{g}_{\text{syn}} H(\psi_6 - \psi_2) + \bar{g}_{\text{syn}} H(\psi_6 - \psi_3).
\end{aligned} \tag{5.16}$$

We first observe that there exist solutions in which the tripods $1, 2, 3$ and $4, 5, 6$ remain internally in-phase. Indeed, seeking (possibly time-dependent) solutions of the form $\psi_1 = \psi_2 = \psi_3 \equiv \psi_L(t)$, $\psi_4 = \psi_5 = \psi_6 \equiv \psi_R(t)$, (5.16) collapses to the pair of equations

$$\dot{\psi}_L = 2\bar{g}_{\text{syn}} H(\psi_L - \psi_R) \quad \text{and} \quad \dot{\psi}_R = 2\bar{g}_{\text{syn}} H(\psi_R - \psi_L), \tag{5.17}$$

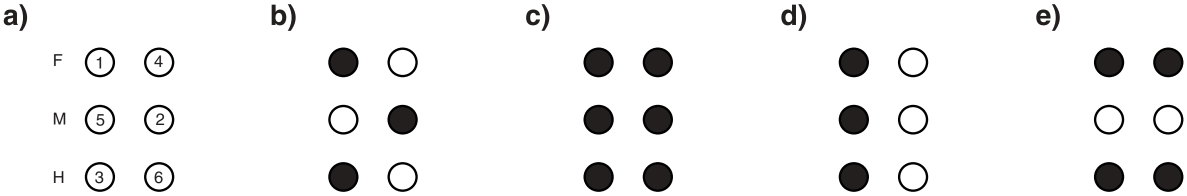


Figure 12. *Some gaits produced by the hexapedal network of Figure* 2(c). (a) *Numbering convention for CPG neurons;* (b) *tripod;* (c) *pronk;* (d) *pace;* (e) *gallop.*

and the arguments used above may be applied to conclude that $\psi_R = \psi_L + \pi$ and $\psi_R = \psi_L$ are fixed points of the $\psi_L - \psi_R$ tripod phase difference equation, again regardless of the form of H. Stability in the full six-dimensional (reduced) phase space is obtained from the 6×6 matrix obtained by linearizing (5.16),

$$\bar{g}_{\text{syn}} \begin{bmatrix} 2H' & 0 & 0 & -H' & -H' & 0 \\ 0 & 2H' & 0 & -H'/2 & -H' & -H'/2 \\ 0 & 0 & 2H' & 0 & -H' & -H' \\ -H' & -H' & 0 & 2H' & 0 & 0 \\ -H'/2 & -H' & -H'/2 & 0 & 2H' & 0 \\ 0 & -H' & -H' & 0 & 0 & 2H' \end{bmatrix}, \tag{5.18}$$

where the derivatives H' are evaluated at the appropriate (constant) phase differences. The antiphase tripod $\psi_L - \psi_R = \pi$ gives one zero eigenvalue with "equal phase" eigenvector $(1, 1, 1, 1, 1, 1)^{\text{T}}$, and the remaining eigenvalues and eigenvectors are as follows:

$$\begin{aligned} &\lambda = \bar{g}_{\text{syn}} H' : \ (1, 0, -1, 1, 0, -1)^{\text{T}}, \\ &\lambda = 2\bar{g}_{\text{syn}} H', \ \text{m} = 2 : \ (1, -1, 1, 0, 0, 0)^{\text{T}}, \text{ and } (0, 0, 0, -1, 1, -1)^{\text{T}}, \\ &\lambda = 3\bar{g}_{\text{syn}} H' : \ (1, 0, -1, -1, 0, 1)^{\text{T}}, \\ &\lambda = 4\bar{g}_{\text{syn}} H' : \ (1, 1, 1, -1, -1, -1)^{\text{T}}. \end{aligned} \tag{5.19}$$

Since $\bar{g}_{\text{syn}} H'(\pi) < 0$ for the nominal parameters, this again indicates asymptotic stability with respect to perturbations that disrupt the tripod phase relationships; moreover, the system recovers fastest from perturbations that disrupt the relative phasing of the tripods ($\lambda = 4\bar{g}_{\text{syn}} H'$: last entry of (5.19)). Since $\bar{g}_{\text{syn}} H'(0) > 0$ (Figure 8(a)), the in-phase "pronking" gait with all legs in phase is unstable.

Other gaits may be found by appealing to discrete symmetries of the network, as in extensive work by Golubitsky and colleagues (see, e.g., [59]). Although they are not directly relevant to cockroach running, we give two examples, shown schematically in Figure 12 along with the antiphase tripod and pronking gaits.

Pace. An appeal to bilateral symmetry $\psi_1 = \psi_5 = \psi_3 = \psi_L$ and $\psi_4 = \psi_2 = \psi_6 = \psi_R$ also yields two equations of the type (5.11), but with an additional term $H(0)$ in each. Since this cancels in subtracting the equations, in- and antiphase solutions again exist.

Gallop. A slightly more complicated gait is obtained by a subgroup of the symmetry group $\mathcal{D}_6$. Setting $\psi_1 = \psi_4 = \psi_F$, $\psi_5 = \psi_2 = \psi_M$, and $\psi_3 = \psi_6 = \psi_H$, (5.16) collapses to the three differential equations

$$
\begin{aligned}
\dot{\psi}_F &= \bar{g}_{\text{syn}}\left[H(0) + H(\psi_F - \psi_M)\right], \\
\dot{\psi}_M &= \bar{g}_{\text{syn}}\left[H(0) + \frac{1}{2}H(\psi_M - \psi_F) + \frac{1}{2}H(\psi_M - \psi_H)\right], \\
\dot{\psi}_H &= \bar{g}_{\text{syn}}\left[H(0) + H(\psi_H - \psi_M)\right].
\end{aligned} \tag{5.20}
$$

This can be further simplified by seeking solutions $\psi_F = \psi_{FH} = \psi_H$ to obtain

$$
\dot{\psi}_{FH} = \bar{g}_{\text{syn}}[H(0) + H(\psi_{FH} - \psi_M)] \quad \text{and} \quad \dot{\psi}_M = \bar{g}_{\text{syn}}[H(0) + H(\psi_M - \psi_{FH})], \tag{5.21}
$$

which is again an instance of (5.11), admitting in-phase and antiphase solutions. The former is just the pronk noted above, but the antiphase "gallop" is new.

5.6. Phase lags and asymmetric coupling. Nominally identical neural oscillators can display differing cycle periods when isolated [60]. Indeed, phase relationships among coupled oscillators can arise from differences in periods as well as from intersegmental coupling characteristics such as strength, projection span, and degree of symmetry [61, 62]. Sensory inputs, moreover, can alter oscillation periods and coordinate mechanical coupling between limbs or segments; they may even form sensory-central oscillatory loops [60]. For example, experimental evidence indicates that the activation of the depressor muscles in *Blaberus discoidalis* does not occur simultaneously even within the tripod gait regime [37]; activation lags are distributed in a range of 0–60% of the cycle. In section 5.3, Figure 8(c), we saw how asymmetric coupling can induce a phase lag, within an antiphase (stable) solution. Here, we extend the two-oscillator analysis to the hexapedal network.

Keeping the left-right symmetry unbroken, we introduce asymmetric ipsilateral coupling $(1-\bar{g}_{\text{F}})$ and $2\bar{g}_{\text{F}}$ between front and middle legs and $(1-\bar{g}_{\text{H}})$ and $2\bar{g}_{\text{H}}$ between hind and middle legs, as shown in Figure 2(d). This leads to modified phase equations

$$
\begin{aligned}
\dot{\psi}_1 &= H(\psi_1 - \psi_4) + 2\bar{g}_{\text{F}}H(\psi_1 - \psi_5), \\
\dot{\psi}_2 &= (1 - \bar{g}_{\text{F}})H(\psi_2 - \psi_4) + H(\psi_2 - \psi_5) + (1 - \bar{g}_{\text{H}})H(\psi_2 - \psi_6), \\
\dot{\psi}_3 &= 2\bar{g}_{\text{H}}H(\psi_3 - \psi_5) + H(\psi_3 - \psi_6), \\
\dot{\psi}_4 &= H(\psi_4 - \psi_1) + 2\bar{g}_{\text{F}}H(\psi_4 - \psi_2), \\
\dot{\psi}_5 &= (1 - \bar{g}_{\text{F}})H(\psi_5 - \psi_1) + H(\psi_5 - \psi_2) + (1 - \bar{g}_{\text{H}})H(\psi_5 - \psi_3), \\
\dot{\psi}_6 &= 2\bar{g}_{\text{H}}H(\psi_6 - \psi_2) + H(\psi_6 - \psi_3),
\end{aligned} \tag{5.22}
$$

where we have included the overall scaling factor $\bar{g}_{\text{syn}}$ in H. We seek solutions which preserve the alternating tripod gait but exhibit phase lags within it:

$$
\begin{aligned}
\psi_4 &= \psi_1 + \pi, \qquad & \psi_5 &= \psi_2 + \pi, \qquad & \psi_6 &= \psi_3 + \pi, \\
\psi_2 &= \psi_1 + \Delta_{\text{F}}, \qquad & \psi_5 &= \psi_4 + \Delta_{\text{F}}, \qquad & \psi_3 &= \psi_2 + \Delta_{\text{H}}.
\end{aligned} \tag{5.23}
$$

(Note that this implies that $\psi_6 = \psi_5 + \Delta_{\text{H}}$, and it automatically ensures phase locking between the tripods: $\dot{\psi}_1 - \dot{\psi}_4 = \dot{\psi}_5 - \dot{\psi}_2 = \dot{\psi}_3 - \dot{\psi}_6 = 0$.) Substituting (5.23) into (5.22) and setting all time derivatives to zero, we obtain two expressions relating the lags $(\Delta_{\rm F}, \Delta_{\rm H})$ to the asymmetry parameters $(\bar{g}_{\rm F}, \bar{g}_{\rm H})$:

$$\begin{aligned} 2\bar{g}_{\rm F} H_{\pi-\Delta_{\rm F}} - (1-\bar{g}_{\rm F}) H_{\pi+\Delta_{\rm F}} - (1-\bar{g}_{\rm H}) H_{\pi-\Delta_{\rm H}} &= 0, \\ \bar{g}_{\rm F} H_{\pi-\Delta_{\rm F}} - \bar{g}_{\rm H} H_{\pi+\Delta_{\rm H}} &= 0. \end{aligned} \tag{5.24}$$

In deriving these, we use 2π-periodicity of H, implying that $H(-\pi+\Delta) = H(\pi+\Delta)$, and we adopt the abbreviated notation $H(\pi \pm \Delta) = H_{\pi\pm\Delta}$. Lacking explicit formulae for H, we cannot solve (5.24) analytically, but rearranging the equations to extract $\bar{g}_{\rm F}, \bar{g}_{\rm H}$,

$$\bar{g}_{\rm F} = \frac{H_{\pi+\Delta_{\rm F}} + H_{\pi-\Delta_{\rm H}}}{\left(2 + \frac{H_{\pi-\Delta_{\rm H}}}{H_{\pi+\Delta_{\rm H}}}\right) H_{\pi-\Delta_{\rm F}} + H_{\pi+\Delta_{\rm F}}}, \quad \bar{g}_{\rm H} = \bar{g}_{\rm F} \frac{H_{\pi-\Delta_{\rm F}}}{H_{\pi+\Delta_{\rm H}}}, \tag{5.25}$$

we can find semiexplicit solutions numerically. Typical slices of the functions $\bar{g}_{\rm F}(\Delta_{\rm F}, \Delta_{\rm H})$ and $\bar{g}_{\rm H}(\Delta_{\rm F}, \Delta_{\rm H})$ are shown in Figures 13(a), (b). Solutions of (5.25) yield the relative synaptic strengths required to achieve given phase lags: for example, Figure 13(a) indicates that setting $\bar{g}_{\rm F} \approx 0.6865$, $\bar{g}_{\rm H} \approx 0.1255$ will give $\Delta_{\rm F} \approx 15^{\rm o}$ $\Delta_{\rm H} \approx 35^{\rm o}$. Figures 13(c), (d) show that for these coupling strengths the equations in (5.22) indeed lock into a tripod gait with $\Delta_{\rm F} = 14.96^{\rm o}$ and $\Delta_{\rm H} = 35.06^{\rm o}$. Finally, Figures 13(e), (f) show that the lags predicted by the phase-reduced theory agree extremely well with those obtained from direct numerical simulations of the full network of (2.1) and (4.1)–(4.2), over a range of biophysically relevant coupling strengths.

5.7. Comparison of phase-reduced and full CPG models. We have already noted (Figures 13(e), (f)) that the phase-reduced model (5.22) and the phase lag/coupling strength relations (5.25) derived from it can predict lags observed in the full network model (2.1), (4.1)–(4.2). We also noted that the symmetric phase-reduction (5.16) correctly captures the stability of the antiphase and instability of of the in-phase solutions for the standard parameter set.

We may go further and observe that the phase-reduced model predicts a timescale for antiphase locking to occur between pairs of oscillators (or, indeed, between left and right tripods). Specifically, linearizing (5.12) at $\theta = \pi$ and using the slope $G'(\pi) \approx -0.15/ms$ from Figure 9(a_4) (solid line), we expect locking to be accomplished within one bursting cycle, and the data of Figures 3, 4, and 5 indicates that this is indeed correct. We also recall that the simple two-oscillator analysis of section 5.3 ((5.9)–(5.10)) predict that pairs of unilaterally coupled CPG and motoneurons should phase lock more readily when the uncoupled frequencies of the former are lower that those of the latter. Figures 14(a)–(d) confirm this with direct simulations of the full network (2.1), (4.1)–(4.2).

The averaged coupling functions obtained in phase reduction also suggest considerable robustness of the (stable) antiphase solution. The multiparametric analyses of section 5.4, illustrated in Figures 9 and 11, show that the slope of G near $\theta = \pi$ remains essentially unchanged even when burst properties are significantly modified (existence of the antiphase fixed point is ensured for any H for the symmetric network of Figure 2(c), as noted in section 5.3). This robustness is implicit in Figure 6, which shows that antiphase tripod solutions of (2.1), (4.1)–(4.2) were found over a substantial domain of a four-dimensional parameter space.

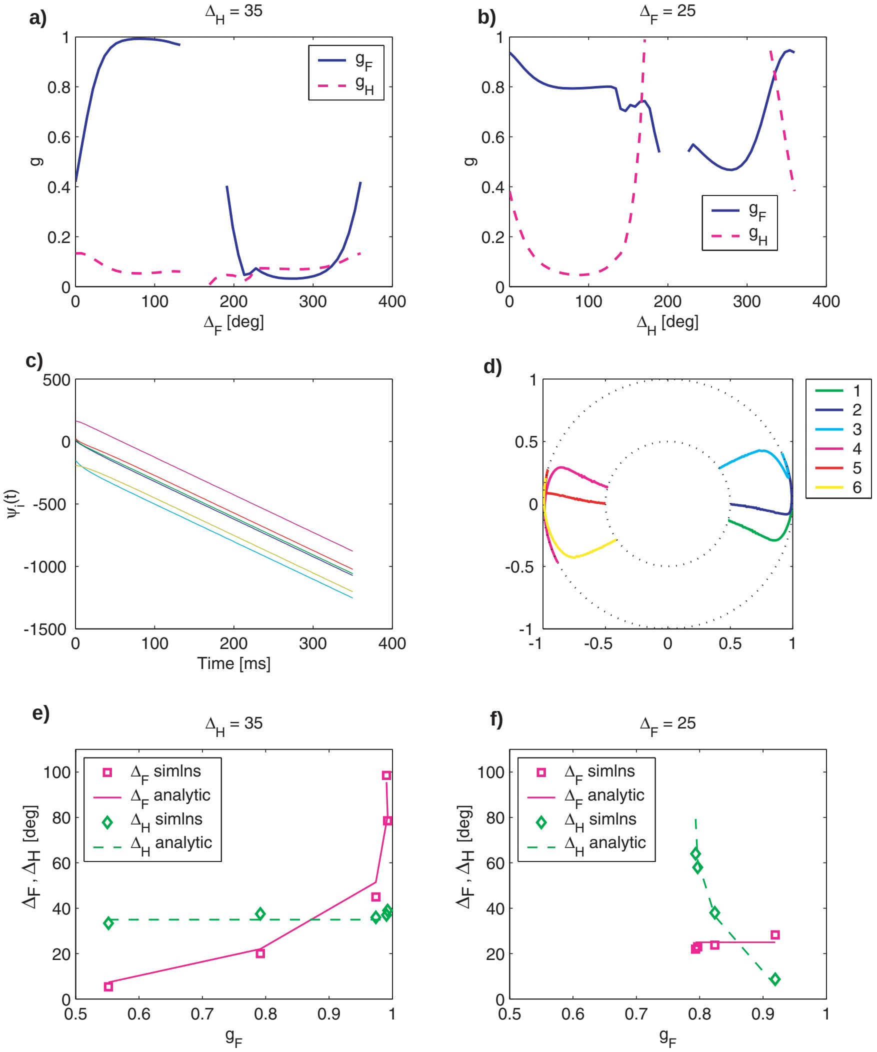


Figure 13. *Phase lagged tripod solutions for the hexapedal network of Figure* 2(d)*. Top panels show* $\bar{g}_F$ *(solid) and* $\bar{g}_H$ *(dashed) as functions of* (a) Δ_F *for fixed* $\Delta_H = 35^o$*, and* (b) Δ_H *for fixed* $\Delta_F = 25^o$*, computed from* (5.25)*. Values* $\bar{g} > 1$ *invert the synapse's "sign" and are invalid, causing broken curves. Panels* (c), (d) *show time histories and polar plots of the six phases with* $\bar{g}_F = 0.6865$, $\bar{g}_H = 0.1255$*, computed from phase-reduced model* (5.22)*. Phases start at* $t = 0$ *on the outer circle and end at* $t = 350$ *msec on the inner circle, having attained the desired lags. Panels* (e), (f) *compare predictions of phase-reduced theory with direct network simulations using* (2.1), (4.1)–(4.2)*: Phase lags* Δ_F *and* Δ_H *from* (5.25) *are shown as solid and dashed lines, and from* (2.1), (4.1)–(4.2) *as squares and diamonds, respectively. In panel* (a), $\bar{g}_F, \bar{g}_H$ *are chosen to keep* Δ_H *fixed, and in* (f), *to keep* Δ_F *fixed.*

Predictions of in-phase solutions with *inhibitory* synapses are even more interesting and potentially delicate. Recalling Figures 9(a_4) and (b_4), we expect stable in-phase solutions for excitatory synapses but *also* for slow inhibitory synapses. That this indeed occurs in the full network is shown in Figures 14(e)–(h): (e) and (f) show anti- and in-phase solutions with excitatory coupling, (g) an in-phase solution which coexists with an antiphase solution, and

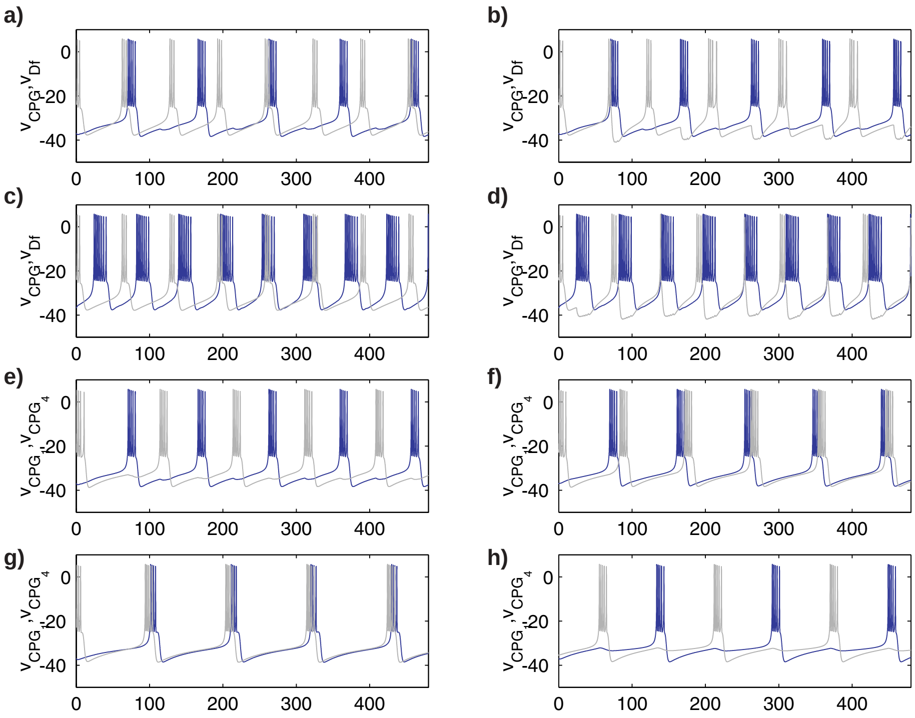


Figure 14. *Membrane voltages of CPG neurons in the hexapedal network of Figure* 2(c)*. Panels* (a), (c) *show uncoupled CPG (dark) and fast motonuerons D_f (grey) with $f_{CPG} = 10.5$ Hz $< f_{D_f} = 15.4$ Hz and $f_{CPG} = 18.1$ Hz $> f_{D_f} = 15.4$ Hz, respectively, and* (b), (d) *show that unidirectional coupling causes* 1:1 *phase locking in the first case, but not the second, which yields* 1:2 *locking. Panels* (e)–(h) *show mutually coupled contralateral CPG neurons* 1 *and* 4*, indicating antiphase locking with inhibitory coupling* (e)*, in-phase locking with excitatory coupling* (f)*, and coexistence of in-phase* (g) *and antiphase* (h) *locking with slow inhibitory coupling, as in Figure* 9(b_4)*;* (g) *and* (h) *are obtained for the same parameter values but different initial conditions. Some panels show the effects of transients, and (in the case of* (f) *and* (g)*) the relatively slow approach to in-phase solutions.*

(h) for slow inhibitory coupling, the latter two solutions being found for identical parameter values but different initial conditions. For such a network both the tripod and the pronk gaits are stable.

6. Conclusions. This paper develops a minimal model for the CPG and representative motoneurons responsible for insect locomotion. We incorporate sufficient biophysical detail to permit appropriate parameter choices and variations to reproduce experimental data, focusing on the cockroaches *Blaberus discoidalis* and *Periplaneta americana*, but we strive for generality and (relative) simplicity. Much current research concerns subcellular details of ionic currents and channels and molecular messengers [63, 64, 65], but despite the ability of "detailed" models to reproduce experimental data (e.g., [66, 7, 67]), their complexity and sensitivity to parameter variations renders them effectively unanalyzable. We believe that massive simulations or experiments alone do not provide global understanding, which profits more from the identification of a few key mechanisms. Thus, our aim is to extract "principles for locomotion" by judicious *selection*, rather than inclusion, of biological data, and in doing so to provide a flexible and tractable mathematical framework within which biological hypotheses can be investigated and novel experiments suggested.

The bursting model (2.1) developed in the preceding paper [1], along with a single equation (4.1) describing synaptic dynamics, is used as the basic subunit to describe the neural architecture of cockroach locomotion. The overall model, which is a "cartoon" representing only a single power stroke (depressor) output per leg, comprises six coupled CPG (inter)neurons, six fast motoneurons, and six slow motoneurons. With appropriate parameter choices, all 18 neurons can be described by the same "minimal" ODE (2.1), and we show how a variety of behaviors, encompassing the range observed in the animals, can be achieved by varying two control parameters separately in the CPG and motoneurons. Since motoneurons are entrained, external currents to CPG interneurons (presumably deriving from higher brain areas and proproceptive feedback) set the stepping frequency, and a CPG conductance primarily determines the duty cycle. Numbers of APs of fast motoneurons and spike rates of slow motoneurons can be separately adjusted by their external currents and conductances, thereby determining muscle forces in coarse and fine manners.

Finally we show how to prove existence, and investigate stability and phase relationships, of gait patterns through an additional reduction using PRCs and averaging theory. This collapses some 60 ODEs of the hexapedal model to six equations for "leg phases" (5.16) and shows that a single network architecture produces a variety of gaits, whose stability properties are primarily determined by the magnitudes and signs of synaptic conductances. We show that the phase-reduced models reproduce the behaviors of the full hexapedal model remarkably well; in particular, Figures 13(e), (f) show phase lags predicted to better than 5% accuracy over a substantial parameter range. It also suggests that further questions regarding how gaits and their stability depend upon neuronal and synaptic parameters will be accessible via the coupling functions H_{ji} of (5.5)–(5.6).

This study, which builds upon earlier work on conservative mechanical and simple actuated models [17, 18, 12, 19, 20], is another step toward integrated neuromechanical models for legged locomotion. In future work we will couple the CPG model developed here to models of muscles and body-limb mechanics and introduce reflexive feedback. In addition to questions on the dynamics and stability of natural gaits such as the double tripod employed by *Blaberus*, and the roles of intrinsic neural parameters and preflexive and reflexive feedback in the CPG, this will allow investigation of questions such as how *Periplaneta* switches to high speed bipedal locomotion [16] and how animals adjust their gaits to quadrupedal patterns within few steps following middle leg amputation [68, pp. 95–99]. Our framework is sufficiently flexible to allow for different numbers of legs and/or motoneurons, for proprioceptive reflexes and CNS feedforward control, as well as for more detailed models of CPG circuitry, and we anticipate that reduced-phase models, with appropriate modifications to PRCs and coupling functions, will continue to provide analytical understanding of such generalized models. Indeed, they hold promise that a CPG model can be coupled to a simple mechanical model to form an integrated neuromechanical system, all in less than 10–15 ODEs.

Acknowledgment. We thank the referees for their close reading and for many valuable suggestions.

REFERENCES

[1] R. Ghigliazza and P. Holmes, *Minimal models of bursting neurons: The effects of multiple currents and timescales*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 636–670.

[2] P. A. Getting, *Comparative analysis of invertebrate central pattern generators*, in Neural Control of Rhythmic Movements in Vertebrates, A. H. Cohen, S. Rossignol, and S. Grillner, eds., John Wiley, New York, 1988, pp. 101–128.

[3] S. Grillner, A. H. Cohen, and S. Rossignol, *Neural Control of Rhythmic Movements in Vertebrates*, Wiley Interscience, New York, 1988.

[4] K. G. Pearson, *Motor systems*, Current Opinion in Neurobiology, 10 (2000), pp. 649–654.

[5] D. M. Wilson, *An approach to the problem of control of rhythmic behaviour*, in Invertebrate Nervous System, C. A. G. Wiersma, ed., The University of Chicago Press, Chicago, IL, 1967, Ch. 17, pp. 219–229.

[6] P. S. Katz, *Intrinsic and extrinsic neuromodulation of motor circuits*, Current Opinion in Neurobiology, 5 (1995), pp. 799–808.

[7] S. Grillner, *Bridging the gap—from ion channels to networks and behaviour*, Current Opinion in Neurobiology, 9 (1999), pp. 663–669.

[8] E. Marder, *Motor pattern generation*, Current Opinion in Neurobiology, 10 (2000), pp. 691–698.

[9] I. E. Brown, S. H. Scott, and G. E. Loeb, *"Preflexes"—programmable, high-gain, zero-delay intrinsic responses of perturbed musculoskeletal systems*, Soc. Neurosci. Abstr., 21 (1995), p. 562.9.

[10] G. E. Loeb, I. E. Brown, and E. J. Cheng, *A hierarchical foundation for models of sensorimotor control*, Exp. Brain Res., 126 (1999), pp. 1–18.

[11] T. M. Kubow and R. J. Full, *The role of the mechanical system in control: A hypothesis of self-stabilization in hexapedal runners*, Phil. Trans. Roy. Soc. London B, 354 (1999), pp. 849–861.

[12] J. Schmitt, M. Garcia, R. Razo, P. Holmes, and R. Full, *Dynamics and stability of legged locomotion in the horizontal plane: A test case using insects*, Biol. Cybern., 86 (2002), pp. 343–353.

[13] D. Jindrich and R. J. Full, *Dynamic stabilization of rapid hexapedal locomotion*, J. Experimental Biology, 205 (2002), pp. 2803–2823.

[14] U. Bässler and A. Büschges, *Pattern generation for stick insect walking movements—multisensory control of a locomotor program*, Brain Research Reviews, 27 (1998), pp. 65–88.

[15] R. J. Full and M. S. Tu, *Mechanics of six-legged runners*, J. Experimental Biology, 148 (1990), pp. 129–146.

[16] R. J. Full and M. S. Tu, *Mechanics of a rapid running insect: Two-, four- and six-legged locomotion*, J. Experimental Biology, 156 (1991), pp. 215–231.

[17] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane—theory*, Biol. Cybern., 83 (2000), pp. 501–515.

[18] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane—application*, Biol. Cybern., 83 (2000), pp. 517–527.

[19] J. Schmitt and P. Holmes, *Mechanical models for insect locomotion: Active muscles and energy losses*, Biol. Cybern., 89 (2003), pp. 43–55.

[20] J. Seipel, P. Holmes, and R. Full, *Dynamics and stability of insect locomotion: A hexapedal model for horizontal plane motions*, Biol. Cybern., 91 (2004), pp. 76–90.

[21] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983; corrected 6th printing, 2002.

[22] C. Morris and H. Lécar, *Voltage oscillations in the barnacle giant muscle*, Biophysics J., 35 (1981), pp. 193–213.

[23] J. Rinzel and G. B. Ermentrout, *Analysis of excitability and oscillations*, in Methods in Neuronal Modeling: From Synapses to Networks, C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1989, pp. 135–169.

[24] L. H. Ting, R. Blickhan, and R. J. Full, *Dynamic and static stability in hexapedal runners*, J. Experimental Biology, 197 (1994), pp. 251–269.

[25] K. G. Pearson, *Discharge patterns of coxal levator and depressor motoneurones in the cockroach, periplaneta americana*, J. Experimental Biology, 52 (1970), pp. 139–165.

[26] K. G. Pearson, *Central programming and reflex control of walking in the cockroach*, J. Experimental Biology, 56 (1972), pp. 173–193.

[27] K. G. Pearson and J. F. Iles, *Nervous mechanisms underlying intersegmental co-ordination of leg movements during walking in the cockroach*, J. Experimental Biology, 58 (1973), pp. 725–744.

[28] K. G. Pearson, *The control of walking*, Scientific American, 464 (1976), pp. 72–86.

[29] K. G. Pearson and C. R. Fourtner, *Nonspiking interneurons in walking system of the cockroach*, J. Neurophysiology, 38 (1975), pp. 33–52.

[30] A. Selverston, R. Elson, M. Rabinovich, R. Huerta, and H. Abarbanel, *Basic principles for generating motor output in the stomatogastric ganglion*, in Neuronal Mechanisms for Generating Locomotor Activity, Vol. 860, O. Kiehn, R. M. Harris-Warrick, N. Kudo, L. M. Jordan, and H. Hultborn, eds., New York Academy of Sciences, New York, 1998, pp. 35–50.

[31] M. Burrows, *The control of sets of motoneurones by local interneurones in the locust*, J. Physiology, 298 (1980), pp. 213–233.

[32] M. Burrows and M. V. S. Siegler, *Networks of local interneurons in an insect*, in Neural Origin of Rhythmic Movements, A. Roberts and B. L. Roberts, eds., Symposia of the Society for Experimental Biology 37, Cambridge University Press, Cambridge, UK, 1983, pp. 29–53.

[33] S. D. Johnston and M.-S. Wu, *Foundations of Cellular Neurophysiology*, MIT Press, Cambridge, MA, 1995.

[34] Y. I. Arshavsky, T. G. Deliagina, and G. N. Orlovsky, *Pattern generation*, Current Opinion in Neurobiology, 7 (1997), pp. 781–789.

[35] J. C. Hancox and R. M. Pitman, *Plateau potentials drive axonal impulse bursts in insect motoneurons*, Proc. Roy. Soc. London Ser. B, Biological Sciences, 244 (1991), pp. 33–38.

[36] A. K. Tryba and R. E. Ritzmann, *Multi-joint coordination during walking and foothold searching in the Blaberus cockroach.* II. *Extensor motor patterns*, J. Neurophysiology, 83 (2000), pp. 3337–3350.

[37] R. J. Full, D. Stokes, A. N. Ahn, and R. K. Josephson, *Energy absorption during running by leg muscles in a cockroach*, J. Experimental Biology, 201 (1998), pp. 997–1012.

[38] A. Longstaff, *Neuroscience*, Instant Notes, Springer-Verlag, New York, 2000.

[39] G. Gradwohl, Y. Grossman, and I. Segev, *Modeling the inhibition of ia input in cat α-motoneurons based on morphologicassl and physiological data*, in Computational Neuroscience. Trends in Research 1995, J. M. Bower, ed., Academic Press, San Diego, 1995, pp. 71–76.

[40] W. Mos, B. L. Roberts, and R. Williamson, *Activity patterns of motoneurons in spinal dogfish in relation to changing fictive locomotion*, Philosophical Transactions of the Royal Society B: Biological Sciences, 330 (1990), pp. 329–339.

[41] C. A. Bishop, J. L. Witten, and M. O'Shea, *Proctolin in the cockroach: Providing model systems for studying neuropeptide transmission*, in Cockroaches as Models for Neurobiology: Applications in Biomedical Research, Vol. 2, I. Huber, E. P. Masler, and B. R. Rao, eds., CRC Press, Boca Raton, FL, 1990, pp. 35–51.

[42] V. Brezina, I. V. Oreknova, and K. R. Weiss, *The neuromuscular transform: The dynamic, nonlinear link between motor neuron firing patterns and muscle contraction in rhythmic behaviors*, J. Neurophysiology, 83 (2000), pp. 207–231.

[43] A. K. Tryba and R. E. Ritzmann, *Multi-joint coordination during walking and foothold searching in the Blaberus cockroach.* I. *Kinematics and electromyograms*, J. Neurophysiology, 83 (2000), pp. 3323–3336.

[44] A. P. Gupta, Y. T. Das, and B. R. Rao, *Anatomy of the cockroach*, in Cockroaches as Models for Neurobiology: Applications in Biomedical Research, Vol. 1, I. Huber, E. P. Masler, and B. R. Rao, eds., CRC Press, Boca Raton, FL, 1990, pp. 33–39.

[45] F. Clarac, *How do sensory and motor signals interact during locomotion?*, in Motor Control: Concepts and Issues, D. R. Humphrey and H.-J. Freund, eds., John Wiley and Sons, Chichester, UK, 1991, pp. 199–221.

[46] S. N. Zill, *Mechanoreceptors: Exteroceptors and proprioceptors*, in Cockroaches as Models for Neurobiology: Applications in Biomedical Research, Vol. 2, I. Huber, E. P. Masler, and B. R. Rao, eds., CRC Press, Boca Raton, FL, 1990, pp. 247–267.

[47] A. Destexhe, Z. F. Mainen, and T. J. Sejnowski, *Kinetic models of synaptic transmission*, in Methods in Neuronal Modeling: From Ions to Networks, 2nd ed., C. Koch and I. Segev, eds., MIT Press, Cambridge, MA, 1999, pp. 1–25.

[48] P. Dayan and L. Abbott, *Theoretical Neuroscience*, MIT Press, Cambridge, MA, 2001.

[49] S. R. Jones, B. Mulloney, T. J. Kaper, and N. Kopell, *Coordination of cellular pattern-generating circuits that control limb movements: The sources of stable differences in intersegmental phases*, J. Neurosci., 23 (2003), pp. 3457–3468.

[50] A. Winfree, *Patterns of phase compromise in biological cycles*, J. Math. Biol., 1 (1974), pp. 73–95.

[51] A. Winfree, *The Geometry of Biological Time*, Springer-Verlag, New York, 2001.

[52] F. C. Hoppensteadt and E. M. Izhikevich, *Weakly Connected Neural Networks*, Appl. Math. Sci. 126, Springer-Verlag, New York, 1980.

[53] G. Ermentrout, *Type* I *membranes, phase resetting curves, and synchrony*, Neural Comp., 8 (1996), pp. 979–1001.

[54] D. Hansel, G. Mato, and C. Meunier, *Phase dynamics for weakly coupled Hodgkin-Huxley neurons*, Europhys. Lett., 23 (1993), pp. 367–372.

[55] J. Guckenheimer, *Isochrons and phaseless sets*, J. Math. Biol., 1 (1975), pp. 259–273.

[56] G. Ermentrout, *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*, Software Environ. Tools 14, SIAM, Philadelphia, 2002.

[57] F. Delcomyn, *Walking robots and the central and peripheral control*, Autonomous Robots, 7 (1999), pp. 259–270.

[58] P. S. G. Stein, M. L. McCullough, and S. N. Currie, *Spinal motor patterns in the turtle*, in Neuronal Mechanism for Generating Locomotor Activity, Vol. 860, O. Kiehn, R. M. Harris-Warrick, L. Jordan, H. Hultborn, and N. Kudo, eds., New York Academy of Sciences, New York, 1998, pp. 51–69.

[59] M. Golubitsky and I. Stewart, *The Symmetry Perspective*, Birkhäuser, Basel, 2002.

[60] W. O. Friesen and J. Cang, *Sensory and central mechanisms control intyersegmental coordination*, Current Opinion in Neurobiology, 11 (2001), pp. 678–683.

[61] E. Marder and R. L. Calabrese, *Principles of rhythmic motor patter generation*, Current Opinion in Neurobiology, 10 (2000), pp. 691–698.

[62] F. K. Skinner and B. Mulloney, *Intersegmental coordination of limb movements during locomotion: Mathematical models predict circuits that drive swimmeret beating*, J. Neurosci., 18 (1998), pp. 3831–3842.

[63] J. Keizer and P. Smolen, *Bursting electrical activity in pancreatic β cells caused by Ca^{2+} and voltage-inactivated Ca^{2+} channels*, Proc. Nat. Acad. Sci. USA, 88 (1991), pp. 3897–3901.

[64] P. Varona, J. J. Torres, R. Huerta, H. D. I. Abarbanel, and M. I. Rabinovich, *Regularization mechanisms of spiking-bursting neurons*, Neural Networks, 14 (2001), pp. 865–875.

[65] R. J. Butera, Jr., J. Rinzel, and J. C. Smith, *Models of respiratory rhythm generation in the pre-Bötzinger complex.* I. *Bursting pacemaker neurons*, J. Neurophysiology, 81 (1999), pp. 382–397.

[66] S. Grillner, P. Wallén, L. Brodin, and A. Lansner, *Neuronal network generating locomotor behavior in lamprey*, Annual Review in Neuroscience, 14 (1991), pp. 169–199.

[67] S. Grillner and P. Wallén, *Cellular basis of a vertebrate locomotor system—steering, intersegmental and segmental co-ordination and sensory control*, Brain Research Review, 40 (2002), pp. 92–106.

[68] D. Graham, *Pattern and control of walking in insects*, in Advances in Insect Physiology, Vol. 18, Academic Press, London, 1985, pp. 31–140.

SIAM J. APPLIED DYNAMICAL SYSTEMS
Vol. 3, No. 4, pp. 701–722

HJB-POD-Based Feedback Design for the Optimal Control of Evolution Problems[*]

K. Kunisch[†], S. Volkwein[†], and L. Xie[†]

Abstract. The numerical realization of closed loop control for distributed parameter systems is still a significant challenge and in fact infeasible unless specific structural techniques are employed. In this paper we propose the combination of model reduction techniques based on proper orthogonal decomposition (POD) with the numerical treatment of the Hamilton–Jacobi–Bellman (HJB) equation for infinite horizon optimal control problems by a modification of an algorithm originated by Gonzales and Rofman and further developed by Falcone and Ferretti. The feasibility of the proposed methodology is demonstrated numerically by means of optimal boundary feedback control for the Burgers equation with noise in the initial condition and in the forcing function.

Key words. dynamic programming, Hamilton–Jacobi–Bellman equation, closed loop control, evolution problems, proper orthogonal decomposition, Burgers equation

AMS subject classifications. 35Kxx, 49Lxx, 65Kxx

DOI. 10.1137/030600485

1. Introduction.

In many applications the discretization of optimal control problems for time dependent partial differential equations, e.g., for the unsteady Navier–Stokes equations, require the solution of nonlinear systems with a large number of degrees of freedom. In particular, to compute closed loop controls in state feedback form we have to solve the Hamilton–Jacobi–Bellman (HJB) equation, which has been numerically infeasible for parabolic differential equations on a standard workstation equipment until today, if classical approximations like finite elements or finite differences are used. In this work model reduction is applied to reduce the number of unknowns significantly. The obtained low-dimensional models should guarantee a reasonable performance of the controlled plant while being computationally tractable. Proper orthogonal decomposition (POD) provides a method for deriving appropriate low-order models. It can be thought of as a Galerkin approximation in the spatial variable, built from functions corresponding to the solution of the physical system at prespecified time instances. These are called the snapshots. Due to possible linear dependence or almost linear dependence a singular value decomposition of the snapshots is carried out and the leading generalized eigenfunctions are chosen as a basis, referred to as the POD basis. Once a low-order model of the dynamical system is available, feedback synthesis based on approximate solutions to the stationary HJB equation becomes feasible.

We demonstrate the feasibility of the proposed approach by means of an optimal boundary control problem for the Burgers equation. Open loop optimal control problems for the

[*] Received by the editors July 18, 2003; accepted for publication (in revised form) by P. Holmes June 17, 2004; published electronically December 27, 2004.

http://www.siam.org/journals/siads/3-4/60048.html

[†] Karl-Franzens-Universität Graz, Institut für Mathematik, Heinrichstrasse 36, A-8010 Graz, Austria (karl.kunisch@uni-graz.at, stefan.volkwein@uni-graz.at, lei.xie@uni-graz.at).

Burgers equation was studied by several authors; see, for instance, [7, 14, 17, 24]. Much less attention has been paid to the important problem of closed loop control. We mention the work by Byrnes, Gilliam, and Shubov [6], where a fixed feedback-operator is used and analyzed, and Burns and Kang [?] where the feedback synthesis is based on Riccati operators for the linearized equations. In [12] instantaneous control was applied to construct a feedback law which matches a desired state, but at considerable control costs. In [15] the authors utilized model reduction with POD to construct a suboptimal feedback synthesis, and an optimal output feedback reduced-order control law was designed by POD discretization in [16].

The analysis and use of proper orthogonal decomposition for reduction purposes has a long-lasting history (see [13] and the references given there). Its use in optimal control, while rather recent, has already created a wide range of literature of which we can only mention a few works. In [19] optimal open loop POD-based control of flow around a rotating cylinder is investigated. Control of turbulent flow utilizing POD with the aim of drag reduction is considered in [20], for example. POD-based control of thin film growth in a chemical vapor deposition reactor is investigated in [3]. In [2] the dynamical system is linearized, which allows the use of Riccati synthesis for feedback controller construction, which can favorably be combined with POD-based model reduction. In [9] and [1] the issue of unmodeled dynamics is addressed, i.e., the fact the snapshots for the POD-approximation are typically taken from dynamics which may be different from the controlled dynamics.

The paper is organized in the following manner: In section 2 we review the dynamic programming principle and the HJB equation. Section 3 is devoted to the reduced-order approach based on POD for an abstract optimal control problem. The numerical strategy for the feedback synthesis is explained in section 4. In section 5 we illustrate the efficiency of the proposed method by considering an optimal boundary control problem for the viscous Burgers equation. Conclusions are drawn in the last section. Some facts about the discretization to the HJB equation that we employ are proven in the appendix.

2. Review of the dynamic programming principle.

In this section we recall the dynamic programming principle and its infinitesimal version, the Hamilton–Jacobi–Bellman equation. This leads to the design of a feedback synthesis by utilizing the so-called value function. For more details we refer the reader to, e.g., [4, Chapter I], [10].

For $k, n \in \mathbb{N}$ let $U = \mathbb{R}^k$ denote the control space and let $U_{\mathsf{ad}} \subsetneq U$ be a closed, bounded, and convex set. Furthermore, $y_\circ \in \mathbb{R}^n$ is a given initial condition. For a measurable control function $u : [0, \infty) \to U$ the state $y : [0, \infty) \to \mathbb{R}^n$ is governed by the initial value problem

$$\dot{y}(t) = F(y(t), u(t)) \qquad \text{for } t > 0, \tag{2.1a}$$

$$y(0) = y_\circ. \tag{2.1b}$$

To ensure the existence of a unique solution to (2.1) we make use of the following assumption.

Assumption 1. *The (nonlinear) mapping $F : \mathbb{R}^n \times U \to \mathbb{R}^n$ is given in such a way that for every choice of initial condition $y_\circ \in \mathbb{R}^n$ and measurable control function u there exists a unique state $y = y(t)$ to the state equation (2.1).*

At times we write $y(t) = y(t; y_\circ, u)$ or $y = y(y_\circ, u)$ to emphasize the dependence of the state y on $y_\circ$ and u. Associated with (2.1) is the cost functional

$$J(y,u) = \int_0^\infty L(y(t),u(t))e^{-\lambda t}\,\mathrm{d}t, \tag{2.2}$$

where $L : \mathbb{R}^n \times \mathbb{R}^k \to [0,\infty)$ is a continuous function and $\lambda > 0$ represents a discount rate.

The optimal control problem is expressed as

$$\min J(y,u) \quad \text{s.t.} \quad y \text{ solves (2.1) and } u \in \mathcal{U}_{\mathsf{ad}}. \tag{2.3}$$

Here, $\mathcal{U}_{\mathsf{ad}}$ denotes the set of all measurable functions from $[0,\infty)$ to U_{ad}. For $u \in \mathcal{U}_{\mathsf{ad}}$ and $y_\circ \in \mathbb{R}^n$ we introduce the reduced cost by

$$\hat{J}(y_\circ,u) = J(y(y_\circ,u),u). \tag{2.4}$$

This gives rise to the value function $v : \mathbb{R}^n \to [0,\infty)$, which is defined by

$$v(y_\circ) = \inf_{u\in\mathcal{U}_{\mathsf{ad}}} \hat{J}(y_\circ,u).$$

It satisfies the *dynamic programming principle*

$$v(y_\circ) = \inf_{u\in\mathcal{U}_{\mathsf{ad}}} \left\{ \int_0^T L(y(t;y_\circ,u),u(t))e^{-\lambda t}\,\mathrm{d}t + v(y(T;y_\circ,u))e^{-\lambda T} \right\} \tag{DPP}$$

for all $y_\circ \in \mathbb{R}^n$ and $T > 0$.

Remark 2.1.

(a) (DPP) holds under general conditions on the data. For example, the existence of optimal control has not been assumed.

(b) When L and, consequently, v are bounded, then $w \equiv v$ holds for every function $w = w(y_\circ)$ satisfying (DPP) for all $y_\circ \in \mathbb{R}^n$ and $T > 0$. ■

Suppose that the value function v is differentiable. Dividing both sides in (DPP) by T and letting T tend to zero, we arrive after a short calculation at the infinitesimal version of the dynamic programming principle, the *HJB equation*:

$$\lambda v(y_\circ) + \sup_{u\in U_{\mathsf{ad}}} \big\{ -\nabla v(y_\circ)F(y_\circ,u) - L(y_\circ,u)\big\} = 0. \tag{HJB}$$

If v is only continuous, then (HJB) has to be interpreted in terms of viscosity solutions. The solution to the HJB equation is utilized for the *synthesis procedure*. Due to the Bellman optimality principle, the function

$$h(t) = v(y^*(t))e^{-\lambda t} + \int_0^t L(y^*(s),u^*(s))e^{-\lambda s}\,\mathrm{d}s$$

is constant for $t > 0$ if and only if $(y^*(y_\circ,u^*),u^*)$ is an optimal trajectory and control pair for the initial condition $y_\circ$. Under the hypothesis that v is differentiable, we conclude that $h' \equiv 0$. In particular, we find

$$\lambda v(y^*(t)) - \nabla v(y^*(t))F(y^*(t),u^*(t)) - L(y^*(t),u^*(t)) = 0 \tag{2.5}$$

for almost all $t > 0$. Utilizing (2.5), it can be shown that under appropriate conditions the control $u^* = u^*(t)$ is optimal if and only if

$$u^*(t) = S(y^*(t)) \quad \text{for almost all } t > 0$$

for any choice S such that

$$S(y_\circ) \in \operatorname*{argmax}_{u \in U_{\mathsf{ad}}} \{ -\nabla v(y_\circ) F(y_\circ, u) - L(y_\circ, u)\}, \tag{2.6}$$

i.e., if and only if

$$\begin{aligned} &\sup_{u \in U_{\mathsf{ad}}} \{ -\nabla v(y^*(t)) F(y^*(t), u) - L(y^*(t), u)\} \\ &\qquad = -\nabla v(y^*(t)) F(y^*(t), u^*(t)) - L(y^*(t), u^*(t)) \quad \text{for almost all } t > 0. \end{aligned}$$

If v was known then determining S would be a finite dimensional mathematical programming problem at every $y_\circ \in \mathbb{R}^n$. S is called the *optimal feedback map.* Assuming that S is known results in the closed loop system

$$\begin{aligned} \dot{y}(t) &= F(y(t), S(y(t))) \quad \text{for } t > 0, \\ y(0) &= y_\circ. \end{aligned} \tag{2.7}$$

Its solution y^* and the optimal control u^* are related by

$$u^*(t) = S(y^*(t)), \quad t > 0. \tag{2.8}$$

We refer to the literature for analogous results if v is only continuous.

For the numerical realization we next discretize (2.1) and (HJB). For the grid size $h > 0$, set

$$t_j = jh \quad \text{for } j = 0, 1, \ldots,$$

and consider the discrete time system

$$\begin{aligned} y_{j+1} &= y_j + hF(y_j, u_j) \quad \text{for } j \geq 0, \\ y_0 &= y_\circ \end{aligned} \tag{2.9}$$

and the associated cost

$$J_h(y_\circ, u_h) = \frac{h}{2} \left(L(y_\circ, u_0) + \sum_{j=1}^{\infty} \beta^j \big(L(y_j, u_{j-1}) + L(y_j, u_j)\big) \right) \tag{2.10}$$

for $u_j \in U_{\mathsf{ad}}$, which arises by applying the trapezoidal rule to (2.2) with the assumption that the controls are constant on the subintervals $[t_{j-1}, t_j]$. Here we set $\beta = e^{-\lambda h}$, and $y_h = \{y_\circ, y_1, \ldots\}$ denotes the solution to (2.9) where $u_h = \{u_0, u_1, \ldots\}$. The approximate minimal value function $v_h : \mathbb{R}^n \to [0, \infty)$ is given by

$$v_h(y_\circ) = \inf_{u_h \in \mathcal{U}_{\mathsf{ad}}^h} J_h(y_\circ, u_h), \tag{2.11}$$

where $\mathcal{U}^h_{\mathsf{ad}} = \{u_h : u_h = \{u_0, u_1, \dots\}$ with $u_i \in U_{\mathsf{ad}}\}$. In the appendix it is verified that v_h is the unique solution to the discrete HJB equation

$$(\mathrm{HJB}_h) \quad v_h(y_\circ) + \sup_{u \in U_{\mathsf{ad}}} \left\{ -\frac{h}{2}\big(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\big) - \beta v(y_\circ + hF(y_\circ, u)) \right\} = 0.$$

Turning to the synthesis problem we define

$$S_h(y_\circ) \in \operatorname*{argmax}_{u \in U_{\mathsf{ad}}} \left\{ -\frac{h}{2}\big(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\big) - \beta v(y_\circ + hF(y_\circ, u)) \right\}.$$

Sufficient conditions, given in the appendix, guarantee that $u^*_j = S_h(y^*_j)$ gives an optimal feedback control, i.e.,

$$v_h(y_\circ) = J_h(y_\circ, u^*)$$

and

$$\begin{aligned} y^*_{j+1} &= y^*_j + hF(y^*_j, S_h(y^*_h)) \quad \text{for } j \geq 0, \\ y^*_0 &= y_\circ. \end{aligned} \tag{2.12}$$

Solving (HJB_h) is still a significant challenge and infeasible for high-dimensional discretizations of distributed parametric systems. For this reason we turn to a model reduction technique in the following section which will allow us to reduce the dimension of the state space y in $\mathbb{R}^n$. The discretization of the value function v_h will be discussed in section 4. We do not address dimension issues concerning the control space U. Certainly, if it is infinite dimensional, it must be discretized for numerical purposes.

3. POD Galerkin approximations for optimal control problems governed by evolution problems.

In this section we propose a reduced-order approach for optimal control problems governed by evolution problems. It is based on POD, which is a method of deriving basis functions containing characteristics of the investigated evolution process. The optimal control problem for an abstract evolution problem and the POD method are introduced in sections 3.1 and 3.2, respectively, and in section 3.3 the reduced-order modeling for the optimal control problem is addressed.

3.1. The optimal control problem for an abstract dynamical system.

Let V and H be real separable Hilbert spaces, and suppose that V is dense in H with compact embedding. By $\langle \cdot\,, \cdot \rangle_H$ we denote the inner product in H. The inner product in V is given by a symmetric bounded, coercive, bilinear form $a : V \times V \to \mathbb{R}$:

$$\langle \varphi, \psi \rangle_V = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V \tag{3.1}$$

with associated norm given by $\| \cdot \|_V = \sqrt{a(\cdot\,, \cdot)}$. We associate with a the linear operator A,

$$\langle A\varphi, \psi \rangle_{V', V} = a(\varphi, \psi) \quad \text{for all } \varphi, \psi \in V,$$

where $\langle \cdot\,, \cdot \rangle_{V', V}$ denotes the duality pairing between V and its dual. Then A is an isomorphism from V onto V'. For $0 < T \leq \infty$ we denote by $L^2(0, T; V)$ the space of equivalence classes of measurable abstract functions $\varphi : (0,T) \to V$, which are square integrable, i.e., $\int_0^T \|\varphi(t)\|_V^2 \,\mathrm{d}t < \infty$. When t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t,\cdot)$ considered as a function in Ω only. The space $W(0,T)$ is defined as

$$W(0,T) = \{\varphi \in L^2(0,T;V) : \varphi_t \in L^2(0,T;V')\},$$

which is a Hilbert space endowed with the common inner product (see, for example, [8, p. 473]), and we set $W_{loc}(0,\infty) = \bigcap_{T>0} W(0,T)$. Let $N : V \to V'$ be a nonlinear continuous operator map. Further, let U be a Hilbert space and $U_{\mathsf{ad}} \subset U$ a closed and convex subset, and set $\mathcal{U} = L^2(0,\infty;U)$ and let $\mathcal{U}_{\mathsf{ad}}$ be the subset of $\mathcal{U}$ containing all functions $u : [0,\infty) \to U_{\mathsf{ad}}$. For $y_\circ \in H$ and $u \in \mathcal{U}_{\mathsf{ad}}$ we consider the nonlinear evolution problem on $[0,\infty)$

$$\frac{\mathrm{d}}{\mathrm{d}t} \langle y(t),\varphi\rangle_H + a(y(t),\varphi) + \langle N(y(t)),\varphi\rangle_{V',V} = \langle B(u(t)),\varphi\rangle_{V',V} \tag{3.2a}$$

for all $\varphi \in V$ and

$$y(0) = y_\circ \quad \text{in } H, \tag{3.2b}$$

where $B : U \to V'$ is a continuous linear operator. We make use of the following assumption.

Assumption 2. *For every $u \in \mathcal{U}_{\mathsf{ad}}$ and $y_\circ \in H$ there exists a unique solution y of* (3.2) *in $W_{loc}(0,\infty)$.*

This assumption is satisfied for many practical situations, including the controlled viscous Burgers and two-dimensional incompressible Navier–Stokes equations.

Next we introduce the cost functional

$$\mathcal{J}(y,u) = \int_0^\infty e^{-\lambda t} \tilde{L}(y(t),u(t))\,\mathrm{d}t,$$

where $\tilde{L} : V \times U \to \mathbb{R}$. The optimal control problem is given by

$$\min \mathcal{J}(y,u) \quad \text{such that} \quad (y,u) \in W_{loc}(0,\infty;V) \times \mathcal{U}_{\mathsf{ad}} \text{ solves (3.2).} \tag{P}$$

Its approximation is considered next.

3.2. The POD method. Throughout we assume that Assumption 2 holds and we denote by y the unique solution to (3.2). For given $n \in \mathbb{N}$ let

$$0 = t_1 < t_2 < \cdots < t_n < \infty \tag{3.3}$$

denote a grid in the interval $[0,\infty)$ and set $\delta t_j = t_j - t_{j-1}$, $j = 1,\ldots,n$. Suppose that the *snapshots* $y_j = y(t_j)$ of (3.2) at the given time instances t_j, $j = 0,\ldots,n$, are known. We set

$$\mathcal{V} = \text{span}\,\{y_0,\ldots,y_n\}.$$

Notice that $\mathcal{V} \subset V$ by construction. Throughout the remainder of this section we let X denote either the space V or the space H.

Let $\{\psi_i\}_{i=1}^d$ denote an orthonormal basis for $\mathcal{V}$ with $d = \dim \mathcal{V}$. Then each member of the ensemble $\mathcal{V}$ can be expressed as

$$y_j = \sum_{i=1}^{d} \langle y_j, \psi_i \rangle_X \psi_i \quad \text{for } j = 0, \ldots, n. \tag{3.4}$$

The method of POD consists in choosing an orthonormal basis such that for every $\ell \in \{1, \ldots, d\}$ the mean square error between the elements y_j, $0 \le j \le n$, and the corresponding ℓth partial sum of (3.4) is minimized on average:

$$\begin{aligned} &\min \mathfrak{J}(\psi_1, \ldots, \psi_\ell) = \sum_{j=0}^{n} \alpha_j \left\| y_j - \sum_{i=1}^{\ell} \langle y_j, \psi_i \rangle_X \psi_i \right\|_X^2 \\ &\text{subject to } \langle \psi_i, \psi_j \rangle_X = \delta_{ij} \quad \text{for } 1 \le i \le \ell, 1 \le j \le i. \end{aligned} \tag{3.5}$$

Here $\{\alpha_j\}_{j=0}^n$ are positive weights, which for our purposes are chosen to be

$$\alpha_0 = \frac{\delta t_1}{2}, \quad \alpha_j = \frac{\delta t_j + \delta t_{j+1}}{2} \text{ for } j = 1, \ldots, n-1, \quad \alpha_n = \frac{\delta t_n}{2}.$$

A solution $\{\psi_i\}_{i=1}^\ell$ to (3.5) is called *POD basis of rank* ℓ. The subspace spanned by the first ℓ POD basis functions is denoted by V^ℓ, i.e.,

$$V^\ell = \text{span}\, \{\psi_1, \ldots, \psi_\ell\}. \tag{3.6}$$

The solution of (3.5) is characterized by the necessary optimality condition, which can be written as an eigenvalue problem. For that purpose we endow $\mathbb{R}^{n+1}$ with the weighted inner product

$$\langle v, w \rangle_{\mathbb{R}^{n+1}} = \sum_{j=0}^{n} \alpha_j v_j w_j \tag{3.7}$$

for $v = (v_0, \ldots, v_n)^T, w = (w_0, \ldots, w_n)^T \in \mathbb{R}^{n+1}$, and the induced norm. Let us introduce the bounded linear operator $\mathcal{Y}_n : \mathbb{R}^{n+1} \to X$ by

$$\mathcal{Y}_n v = \sum_{j=0}^{n} \alpha_j v_j y_j \quad \text{for } v \in \mathbb{R}^{n+1}. \tag{3.8}$$

Then the adjoint $\mathcal{Y}_n^* : X \to \mathbb{R}^{n+1}$ is given by

$$\mathcal{Y}_n^* z = \left(\langle z, y_0 \rangle_X, \ldots, \langle z, y_n \rangle_X \right)^T \quad \text{for } z \in X. \tag{3.9}$$

It follows that $\mathcal{R}_n = \mathcal{Y}_n \mathcal{Y}_n^* \in \mathcal{L}(X)$ and $\mathcal{K}_n = \mathcal{Y}_n^* \mathcal{Y}_n \in \mathbb{R}^{(n+1)\times(n+1)}$ are given by

$$\begin{aligned} \mathcal{R}_n z &= \sum_{j=0}^{n} \alpha_j \langle z, y_j \rangle_X y_j \quad \text{for } z \in X, \\ \left(\mathcal{K}_n\right)_{ij} &= \alpha_j \langle y_j, y_i \rangle_X, \end{aligned} \tag{3.10}$$

respectively. Here, $\mathcal{L}(X)$ denotes the Banach space of all bounded linear operators from X into itself. The matrix $\mathcal{K}_n$ is often called a *correlation matrix*.

Using a Lagrangian framework, we can derive the following optimality conditions for the optimization problem (3.5):

$$\mathcal{R}_n\psi = \lambda\psi \tag{3.11}$$

(see, e.g., [13, pp. 88–91] and [23, section 2]). Note that $\mathcal{R}_n$ is a bounded, self-adjoint, and nonnegative operator. Moreover, since the image of $\mathcal{R}_n$ is finite dimensional, $\mathcal{R}_n$ is also compact. By Hilbert–Schmidt theory (see, e.g., [21, p. 203]) there exist an orthonormal basis $\{\psi_i\}_{i=1}^\infty$ for X and a sequence $\{\lambda_i\}_{i=1}^\infty$ of nonnegative real numbers so that

$$\mathcal{R}_n\psi_i = \lambda_i\psi_i, \quad \lambda_1 \geq \cdots \geq \lambda_d > 0, \quad \text{and } \lambda_i = 0 \quad \text{for} \quad i > d. \tag{3.12}$$

Moreover, $\mathcal{V} = \operatorname{span}\,\{\psi_i\}_{i=1}^d$. Note that $\{\lambda_i\}_{i=0}^\infty$ as well as $\{\psi_i\}_{i=0}^\infty$ depend on n. Contents permitting the notation of this dependence is dropped.

Remark 3.1. Setting

$$v_i = \frac{1}{\sqrt{\lambda_i}}\,\mathcal{Y}_n^*\psi_i \quad \text{for } i = 1, \ldots, d,$$

we find $\mathcal{K}_n v_i = \lambda_i v_i$ and $\langle v_i, v_j\rangle_{\mathbb{R}^{n+1}} = \delta_{ij}$ for $1 \leq i, j \leq d$. Thus, $\{v_i\}_{i=1}^d$ is an orthonormal basis of eigenvectors of $\mathcal{K}_n$ for the image of $\mathcal{K}_n$. Conversely, if $\{v_i\}_{i=1}^d$ is a given orthonormal basis for the image of $\mathcal{K}_n$, then it follows that the first d eigenfunctions of $\mathcal{R}_n$ can be determined by

$$\psi_i = \frac{1}{\sqrt{\lambda_i}}\,\mathcal{Y}_n v_i \quad \text{for } i = 1, \ldots, d.$$

Hence, we can determine the POD basis by solving either the eigenvalue problem for $\mathcal{R}_n$ or the one for $\mathcal{K}_n$. ■

The sequence $\{\psi_i\}_{i=1}^\ell$ solves the optimization problem (3.5). This fact as well as the error formula below were proved in [13, section 3], for example. Let $\lambda_1 \geq \cdots \geq \lambda_d > 0$ denote the positive eigenvalues of $\mathcal{R}_n$ with the associated eigenvectors $\psi_1, \ldots, \psi_d \in X$. Then, $\{\psi_i^n\}_{i=1}^\ell$ is a POD basis of rank $\ell \leq d$, and we have the error formula

$$\mathfrak{J}(\psi_1, \ldots, \psi_\ell) = \sum_{j=0}^n \alpha_j \left\| y_j - \sum_{i=1}^\ell \langle y_j, \psi_i\rangle_X \psi_i \right\|_X^2 = \sum_{i=\ell+1}^d \lambda_i. \tag{3.13}$$

3.3. Reduced-order control. The reduced-order approach to optimal control problems such as (P) is based on approximating the nonlinear dynamics by a Galerkin technique utilizing basis functions that contain characteristics of the controlled dynamic.

To compute a POD solution of (P) we make the ansatz

$$y^\ell(t, x) = \sum_{i=1}^\ell \mathrm{w}_i(t)\psi_i(x). \tag{3.14}$$

We introduce the mass and stiffness matrices by

$$\begin{aligned}\mathrm{M} &= ((\mathrm{m}_{ij})) \in \mathbb{R}^{\ell\times\ell} \text{ with } \mathrm{m}_{ij} = \langle \psi_j, \psi_i\rangle_H,\\ \mathrm{S} &= ((\mathrm{s}_{ij})) \in \mathbb{R}^{\ell\times\ell} \text{ with } \mathrm{s}_{ij} = a(\psi_j, \psi_i),\end{aligned}$$

the nonlinear function $\mathrm{N} : \mathbb{R}^\ell \to \mathbb{R}^\ell$ by

$$(\mathrm{w}_1, \dots, \mathrm{w}_\ell) \mapsto \mathrm{N}(\mathrm{w}_1, \dots, \mathrm{w}_\ell) = (\mathrm{n}_i) \in \mathbb{R}^\ell \text{ with } \mathrm{n}_i = \left\langle N\left(\sum_{j=1}^{\ell} \mathrm{w}_i \psi_j\right), \psi_i \right\rangle_{V',V},$$

and the mapping of the control input $\mathrm{b} : U \to \mathbb{R}^\ell$ by

$$u \mapsto \mathrm{b}(u) = \big(\mathrm{b}(u)_i\big) \in \mathbb{R}^\ell \text{ with } \mathrm{b}(u)_i = \langle Bu, \psi_i\rangle_H.$$

The modal coefficients of the initial condition $y^\ell(0) \in \mathbb{R}^\ell$ are determined by $\mathrm{w}_i(0) = (\mathrm{w}_\circ)_i = \langle y_\circ, \psi_i\rangle_X$, $1 \le i \le \ell$, and the solution vector of the reduced dynamical system is denoted by $\mathrm{w}^\ell(t) \in \mathbb{R}^\ell$. Then the Galerkin approximation of the optimal control problem (P) is given by

$$(\mathrm{P}^\ell) \qquad \begin{cases} \min J^\ell(\mathrm{w}^\ell, u) \\ \text{s.t. } u \in \mathcal{U}_{\mathsf{ad}} \text{ and } \begin{cases} \dot{\mathrm{w}}^\ell(t) = F(\mathrm{w}^\ell(t), u(t)) & \text{for } t > 0, \\ \mathrm{w}^\ell(0) = \mathrm{w}_\circ, \end{cases} \end{cases}$$

where the cost functional is defined as

$$J^\ell(\mathrm{w}^\ell, u) = \int_0^\infty \tilde{L}(y^\ell(t), u(t)) e^{-\lambda t}\,\mathrm{d}t$$

with w^ℓ and y^ℓ related by (3.14) and the nonlinear mapping $F : \mathbb{R}^\ell \times U \to \mathbb{R}^\ell$ given by

$$F(\mathrm{w}^\ell, u) = \mathrm{M}^{-1}\left(-\mathrm{S}\mathrm{w}^\ell - \mathrm{N}(\mathrm{w}^\ell) + \mathrm{b}(u)\right).$$

Of course, it is tacitly assumed that the dynamical system in (P^ℓ) admits a unique solution for every $u \in \mathcal{U}_{\mathsf{ad}}$. Let us mention that in case of $X = H$ the mass matrix M is just the identity matrix. On the other hand, S is the identity matrix for $X = V$.

The value function v^ℓ, defined for initial states $\mathrm{w}_\circ \in \mathbb{R}^\ell$, is

$$v^\ell(\mathrm{w}_\circ) = \inf_{u \in \mathcal{U}_{\mathsf{ad}}} \hat{J}^\ell(\mathrm{w}_\circ, u),$$

where $\hat{J}^\ell(\mathrm{w}_\circ, u) = J^\ell(\mathrm{w}^\ell, u)$ and w^ℓ solves the dynamical system in (P^ℓ) with control input u and initial condition $\mathrm{w}_\circ$.

4. Numerical strategy for the closed loop design.

Here we briefly explain the numerical realization of (HJB_h). While (HJB_h) is defined on $\mathbb{R}^n$ for practical purposes, we restrict ourselves to a computational domain Υ_h which is a bounded subset of $\mathbb{R}^n$. This is justified if $y + hF(y, u) \in \Upsilon_h$ for all $y \in \Upsilon_h$ and $u \in U_{\mathsf{ad}}$. Here we choose $\Upsilon_h = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_\ell, b_\ell]$ with $a_1 \ge a_2 \ge \cdots \ge a_\ell$ and $b_1 \ge b_2 \ge \cdots \ge b_\ell$. Let $\{S_j\}_{j=1}^k$ denote the hypercubes of a rectilinear partition of Υ_h with N vertices $\{y_j\}$. We consider the space W^k of piecewise ℓ-linear functions $w^k : \Upsilon_h \to \mathbb{R}$ which are continuous on Υ_h. We look for a solution $w^k \in W^k$ of

$$(\mathrm{HJB}_h^k) \qquad w^k(y_j) = \inf_{u \in U_{\mathsf{ad}}} \left\{ \frac{h}{2}(L(y_j,u) + \beta L(y_j + hF(y_j,u),u)) + \beta w^k(y_j + hF(y_j,u)) \right\}$$

for every vertex $y_j \in \Upsilon_h$. If v_h is a solution to (HJB_h), then it satisfies (HJB_h^k) and the convex interpolant

$$w^k(y) = \sum_{j=1}^{N} \lambda_j v(y_j) \text{ for } y = \sum_{j=1}^{N} \lambda_j y_j$$

belongs to W^k. The intervals $[a_j, b_j]$ must be chosen such that they contain the components of the expected controlled trajectories while simultaneously keeping them as small as possible for computational purposes. Since the dynamic model is derived from POD, the magnitude of the components in the solution representation of the trajectory in terms of the POD basis functions are rapidly decreasing. We therefore choose the intervals $[a_j, b_j]$ such that their lengths are rapidly decreasing. Moreover, the mesh sizes decrease as j increases, since we expect the solution to be more sensitive to modes with lower index. The evaluation of the right-hand side of (HJB_h^k) requires us to solve a constrained nonlinear programming problem. For this purpose we used `fmincon` from Matlab. It would certainly be worthwhile to investigate possible speed-up by employing different techniques for this task. To solve (HJB_h^k) on Υ_h, a fixed point iteration with a multilevel acceleration strategy was used. Once (HJB_h^k) is solved, the optimal value function and corresponding optimal control at each grid point are available.

1. Multilevel method. The convergence rate of the fixed point iteration depends on h. The HJB equation is first solved for $h = 0.2$; the result is taken as initial guess to solve the HJB equation for $h = 0.05$. These two results are utilized to predict a new guess by means of the secant method. Then the HJB equation is solved for $h = 0.0125$. This multilevel method significantly accelerates the fixed point iteration.

2. Parallel computation. In solving the dynamic programming equation, the constrained minimizing problem must be solved at each point of the polyhedron; these computations are independent of each other and can therefore be performed in a fine-grained parallel strategy. In this parallelism, the same set of codes runs simultaneously on different pieces of data on various processors. The technique of *massage passing interface* (MPI) [11] was used in our numerical tests. The mechanism used in MPI to distribute data (or information) is through explicit sending and receiving of data among the processors. The newest MPI standard was released in 1997. We refer to [18].

5. Application to the viscous Burgers equation.

In this section we demonstrate the efficiency of the proposed methodology by means of optimal boundary control of the viscous Burgers equation.

5.1. The optimal control problem.

Define the domains $\Omega = (0,1) \subset \mathbb{R}$, $Q = (0,\infty) \times \Omega$, and $\Sigma = (0,\infty) \times \partial\Omega$. In the context of section 3.1, we set $H = L^2(\Omega)$, $V = H^1(\Omega)$, and we define

$$a(\varphi, \phi) = \nu \int_\Omega \varphi' \phi' \,\mathrm{d}x \quad \text{for } \varphi, \phi \in V,$$

with $\nu > 0$ and $B \in \mathcal{L}(\mathbb{R}, V')$ by

$$\langle Bu, \phi \rangle_{V',V} = u\,\phi(0).$$

Let $u_a \leq u_b$. We set $U_{\mathsf{ad}} = \{u \in \mathbb{R} : u_a \leq u \leq u_b\}$ and define the set of admissible controls

$$\mathcal{U}_{\mathsf{ad}} = \{u \in L^2_{loc}(0,\infty) : u(t) \in U_{\mathsf{ad}} \text{ for almost all } t \in (0,\infty)\}. \tag{5.1}$$

For a control $u \in \mathcal{U}_{\mathsf{ad}}$ we consider the viscous Burgers equation

$$y_t - \nu y_{xx} + y y_x = 0 \qquad \text{in } Q, \tag{5.2a}$$

$$\nu y_x(\cdot,0) + \sigma_0 y(\cdot,0) = u \qquad \text{in } (0,\infty), \tag{5.2b}$$

$$\nu y_x(\cdot,1) + \sigma_1 y(\cdot,1) = g \qquad \text{in } (0,\infty), \tag{5.2c}$$

$$y(0,\cdot) = y_\circ \qquad \text{in } \Omega, \tag{5.2d}$$

where $y_\circ \in L^2(\Omega)$ is a given initial condition and σ_0, σ_1, and g are real numbers. Henceforth we consider weak solutions $y \in W_{loc}(0,\infty;V)$ of (5.2) satisfying (5.2d) and

$$\begin{aligned} &\langle y_t(t), \varphi \rangle_{V',V} + \sigma_1 y(t,1)\varphi(1) - \sigma_0 y(t,0)\varphi(0) \\ &a(y,\varphi) + \int_\Omega y(t) y'(t) \varphi \,\mathrm{d}x = g\varphi(1) - \langle Bu, \varphi \rangle_{V',V} \end{aligned} \tag{5.3}$$

for all $\varphi \in H^1(\Omega)$ and $t \in (0,\infty)$ a.e. For the functional analytic treatment of (5.2) we refer to [22, 24], for example. We shall consider the cost functional

$$J(y,u) = \int_0^\infty \left(\frac{1}{2} \int_\Omega |y(t,x) - z(x)|^2 \,\mathrm{d}x + \frac{\beta}{2}\, |u(t)|^2 \right) e^{-\lambda t} \,\mathrm{d}t,$$

where $z \in L^2(\Omega)$ is a given desired state and $\lambda, \beta > 0$ are positive constants.

The optimal control problem is given by

$$\min J(y,u) \quad \text{such that} \quad (y,u) \in W_{loc}(0,\infty) \times \mathcal{U}_{\mathsf{ad}} \text{ satisfies (5.2)}, \tag{$\tilde{\mathrm{P}}$}$$

as a weak solution. It is straightforward to argue the existence of an optimal control for ($\tilde{\mathrm{P}}$).

5.2. Reduced-order control. Suppose that we have computed a POD basis utilizing, e.g., a finite element code for the viscous Burgers equation and determined the basis functions as described in section 3.2. To compute a POD solution of ($\tilde{\mathrm{P}}$) we make the ansatz (3.14) for the state variable. In addition to the matrices and vectors defined in section 3.3 we introduce the tensor

$$\mathrm{T} = (((\mathrm{b}_{ijk}))) \in \mathbb{R}^{\ell \times \ell \times \ell} \text{ with } \mathrm{b}_{ijk} = \int_\Omega \psi_j \psi_k' \psi_i \,\mathrm{d}x,$$

and the vectors for the boundary conditions

$$\mathrm{d} = (\mathrm{d}_i) \in \mathbb{R}^\ell \text{ with } \mathrm{d}_i = \psi_i(0), \quad \mathrm{e} = (\mathrm{e}_i) \in \mathbb{R}^\ell \text{ with } \mathrm{e}_i = \psi_i(1).$$

Table 1
Construction of parallel computations.

Node	Portion
master	12/18
slave1	3/18
slave2	2/18
slave3	1/18

Then the Galerkin approximation of the optimal control problem $(\tilde{\mathrm{P}})$ is given by

$$(\tilde{\mathrm{P}}^\ell) \qquad \begin{cases} \min J^\ell(\mathrm{w}^\ell, u) \\ \text{s.t. } u \in \mathcal{U}_{\mathsf{ad}} \text{ and } \begin{cases} \dot{\mathrm{w}}^\ell(t) = F(\mathrm{w}^\ell(t), u^\ell(t)) & \text{for } t > 0, \\ \mathrm{w}^\ell(0) = \mathrm{w}_\circ, \end{cases} \end{cases}$$

where the nonlinear mapping $F : \mathbb{R}^\ell \times \mathbb{R} \to \mathbb{R}^\ell$ is defined by

$$F(\mathrm{w}^\ell, u) = \mathrm{M}^{-1}\Big((- \mathrm{S} - (\mathrm{T} : \mathrm{w}^\ell))\mathrm{w}^\ell + \mathrm{d}\left(\mathrm{d}^\mathsf{T}\sigma_0\mathrm{w}^\ell - u\right) - \mathrm{e}(\mathrm{e}^\mathsf{T}\sigma_1\mathrm{w}^\ell - g)\Big).$$

The value function v, defined for any initial state $\mathrm{w}_\circ \in \mathbb{R}^\ell$, is

$$v(\mathrm{w}_\circ) = \inf_{u \in \mathcal{U}_{\mathsf{ad}}} \hat{J}^\ell(\mathrm{w}_\circ, u),$$

where $\hat{J}^\ell(\mathrm{w}_\circ, u) = J^\ell(\mathrm{w}^\ell, u)$ and w^ℓ solves the dynamical system in $(\tilde{\mathrm{P}}^\ell)$ with initial condition $\mathrm{w}_\circ$ and control input u.

5.3. Numerical experiments. This subsection is devoted to demonstrate the efficiency of the feedback synthesis proposed in section 4.

In practical implementations, three Matlab sessions are started on three slaves remotely from the master. Then the required data are transferred to the slaves via MPI. On receiving data, each slave can perform computations concurrently. The portion of the computational work to be performed on each slave can be adjusted according to the performance of the slaves. After all computations are done on the slaves, the data will be collected from the slaves. The distribution of the parallel computation is shown in Table 1. The consumed time (in seconds) are displayed in Table 2 for the parallel and serial computations to calculate one iteration of the fixed point scheme. The specific numbers correspond to the example with discontinuous initial data, given below. The last row shows the ratio of the parallel time cost to the serial time cost. With the number of grid points increasing, the ratio is increasing, partly because more time is consumed to transfer required data to and from the slaves.

Two computational tests will be presented, one with continuous initial condition and the other with discontinuous initial condition. For the sake of comparison we also compute open loop solutions. This can be done efficiently by means of SQP techniques applied to $(\tilde{\mathrm{P}})$ [24], where the constraint in the form of the Burgers equation is discretized by a finite element technique. Moreover, the infinite time horizon was replaced by a finite horizon $[0, T]$, with T chosen sufficiently large so that it has little effect on the numerical results. The parameter

Table 2

Comparisons of parallel and serial computations: CPU times in seconds.

Grid points	625	5525	9945	17901
Parallel	20.94	194.43	438.86	814.90
Serial	35.09	344.60	643.37	1094.16
	59.68%	56.42%	68.21%	74.75%

Table 3

Parameter settings.

Symbol	Value	Description
λ	2.0	discount rate
β	0.05	weighting coefficient for the control
T	5	time horizon
z	0	desired state

settings are listed in Table 3. Concerning the boundary conditions for the Burgers equation (5.2), we set $\sigma_0 = \sigma_1 = 0, g = 0$. We took 251 equidistant snapshots from the uncontrolled dynamics. For both examples four basis functions are used for the POD approximation. In terms of the ratio $r(\ell) = \sum_1^\ell \lambda_i / \sum_{i=\ell+1}^d \lambda_i$ this means that $r(4) \geq .985$ for the first example below, and $r(4) \geq .9999$ for the second example. Unless specified otherwise, the grid size was chosen to be $24 \times 16 \times 4 \times 4$. We also report on the effect of the choice of this grid. In our numerical tests we frequently replaced the explicit Euler approximation $y_\circ + hF(y_\circ, u)$ of $y(h)$ by a semiimplicit approximation of $y(h)$. This improved the performance without qualitatively changing the results. Finally, let us comment on the choice of snapshots, which were taken from the uncontrolled dynamics for the results to be presented below. We also carried out tests with taking snapshots from the dynamics, controlled by the open loop optimal control, and combination of the former and the latter. There was little effect on the value of the cost J (evaluated for the closed loop optimal control and the associated trajectory). However, the difference between this value for J and the value of the value-function obtained from the HJB equation, which, as we explain below, is used for validation of our procedure, increases. This comes as no surprise for the class of test problems under consideration. In fact, the controlled states converge to the origin rather quickly and hence contain significantly less information than the uncontrolled snapshots resulting in a decrease of the approximation property of the HJB equation. This in turn could possibly be counteracted by taking nonuniformly spaced snapshots, an issue that we do not want to pursue in this work.

Continuous initial condition. In this case the continuous initial condition is $y(0) = (1-x)\sin(3\pi(x-0.5))$, and the viscosity coefficient is $\nu = 0.05$. The state evolutions without control and with feedback optimal control are displayed in Figure 1. As expected, the controlled state decreases as time evolves. The feedback and open loop controls are compared in Figure 2. In Table 4, we can see that the cost functional is decreased from 0.01818 to 0.00766 in the feedback design and from 0.01812 to 0.00681 in the open loop design. This minor difference is not unexpected since the feedback design is based on the reduced system obtained by the POD technique, whereas the open loop optimal control is computed by means of an

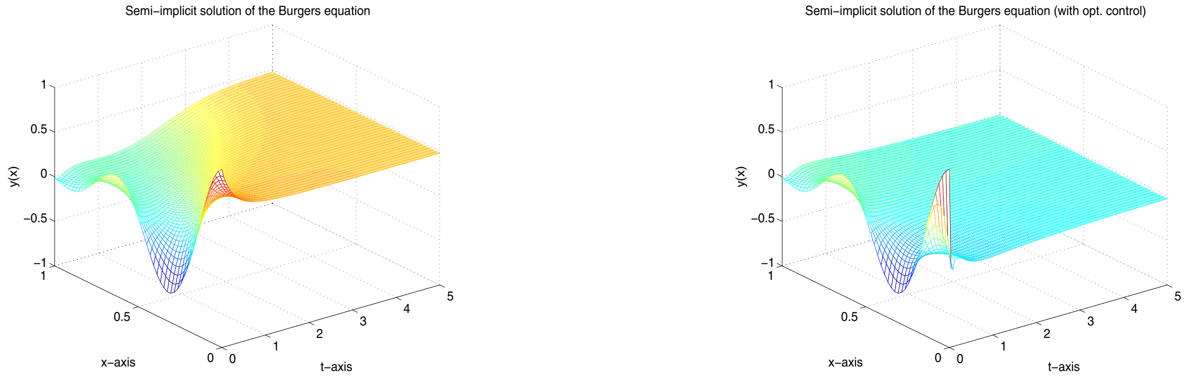


Figure 1. *Uncontrolled state (left) and optimal state (right): Continuous initial condition.*

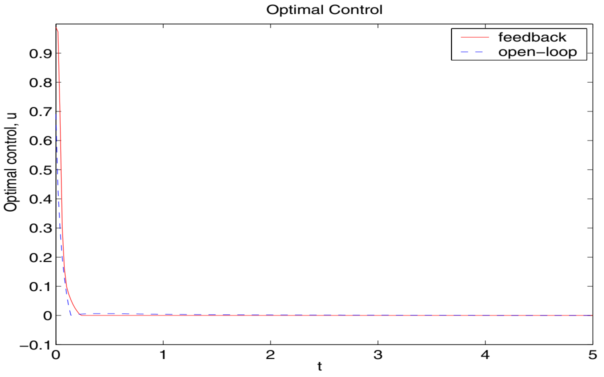


Figure 2. *Comparisons of optimal controls from feedback and open-loop design: Continuous initial condition.*

SQP technique for a high-resolution finite element discretization of a continuous system (5.2). A further validation of the numerical results is obtained by comparing the values of the cost function obtained from (i) the open loop control as in the second row of the right column of Table 4 , (ii) inserting the controls and the controlled state into the cost as in the second row of the left column, and (iii) from the numerical approximation to the HJB equation, shown in the first column's last row.

Discontinuous initial condition. In this test case, the initial condition is

$$y_\circ(x) = \begin{cases} 1 & \text{if } 0 \le x < 0.5, \\ 0 & \text{if } 0.5 < x \le 1, \end{cases} \tag{5.4}$$

and the viscosity coefficient is $\nu = 0.25$.

First we carry out a grid convergence study using $\lambda = 1$ with all other specifications as in Table 3. The results are shown in Table 5. Note, in particular, that the difference between J and V decreases as the grid is refined. In the following tests, we will take the grid system of $24 \times 16 \times 4 \times 4$ for the polyhedron, which was also taken for the case with the continuous initial condition.

Table 4

Comparisons of results from feedback and open loop design: Continuous initial condition.

	Feedback	Open loop
J w opt. control	0.00766	0.00681
J w/o control	0.01818	0.01812
Value function	0.00735	

Table 5

Grid convergence study: The difference between optimal cost functional and value function decreases as the grid of the polyhedron is fined.

	$4 \times 4 \times 4 \times 4$	$16 \times 12 \times 4 \times 4$	$24 \times 16 \times 4 \times 4$	$32 \times 16 \times 5 \times 5$
J w opt. control	0.0320	0.0319	0.0319	0.0318
J w/o control	0.1267	0.1267	0.1267	0.1267
Value function	0.0272	0.0354	0.0336	0.0324
Comp. time (units)	0.08	1	2	6
Error: V and J	-15%	11 %	5 %	2%

The evolutions of the state are depicted in Figure 3 for the uncontrolled and controlled cases. Furthermore, as observed in the discussion of the results in Figure 4, the feedback control agrees well with the open loop design result. The computational results are summarized in Table 6. Again we can claim good agreement between the optimal cost functional and the value functional based on the reduced order calculations.

Let us turn to the effect of noise. First random noise is imposed on the initial condition. The open loop design fails to drive the system to zero, if uniform noise in $[-9, 9]$ is added to the initial condition. The feedback design, however, can still generate an acceptable result, as shown in Figure 5. Another test considered here is to impose random noise on the right-hand side of the Burgers equation (5.2a). The controlled states with random uniform random noise in $[-0.25, 0.25]$ (constant w.r.t. t) are displayed in Figure 6, respectively, for feedback and open loop design. Comparing the controlled states at $t = 5$ the feedback result is clearly better than the open loop one. The reader will note a drift in the controlled solution, to a value below 0, for the specific realization of the random numbers for this numerical run. Let us point out here the behavior of the uncontrolled Burgers equation with Neumann boundary conditions and random forcing with zero mean: the solution tends to be constant w.r.t. x with the constant depending on the mean of the concrete realization of the set of random numbers (which happens to be negative for the numerical example depicted in Figure 6).

These comparisons confirm that the reduced-order HJB-based closed loop control design is effective in the presence of noise in the system dynamics.

6. Conclusion.

This paper deals with nonlinear feedback design for evolution problems. The feedback gain is obtained as the solution of the discrete HJB equation. Since the spatial dimension for the HJB equation depends on the number of spatial grid points used in the numerical scheme for the evolution problem, the size of the HJB equation is numerically infeasible if, e.g., finite element or finite difference approximations are used. Here reduced-order modeling with POD is applied for the spatial discretization of the dynamical system resulting in a low-dimensional HJB equation, which can be solved by a fixed-point–type algorithm. To

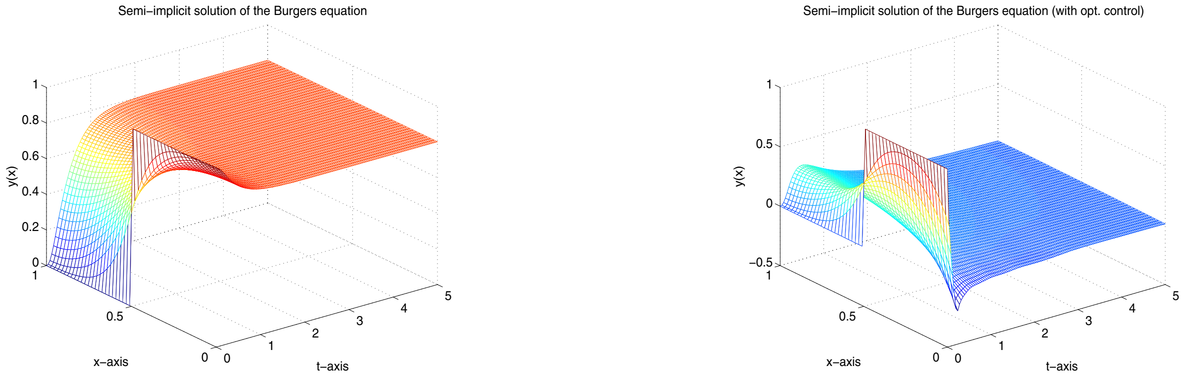


Figure 3. *Uncontrolled state (left) and optimal state (right): Discontinuous initial condition.*

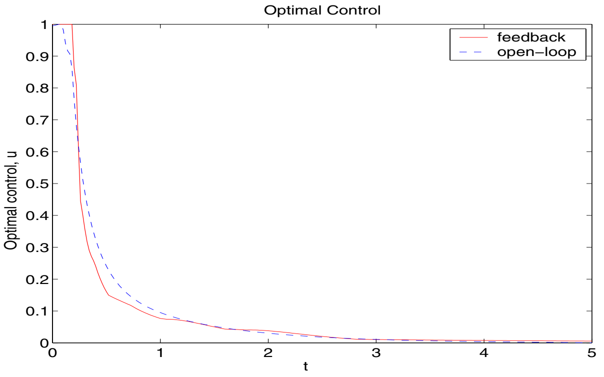


Figure 4. *Comparison of optimal control from feedback and open loop design: Discontinuous initial condition.*

accelerate the method both nested iterations and parallelization are utilized. The numerical strategy is illustrated numerically by taking an optimal boundary control problem for the Burgers equation. It turns out that the closed loop control can be computed with reasonable effort. Moreover, the feasibility of the proposed method and the superiority to open loop control is demonstrated by examples including noise in the initial condition and in the forcing function.

Appendix.

Here we verify the claims made in the second part of section 2. Throughout we assume that $h \in (0,1]$ and that there exist constants M, L_1, L_2 such that

$$\begin{aligned}
&\text{(A.1)} \qquad & hL(y,u) \le M \qquad & \text{for all } (y,u) \in \mathbb{R}^n \times U_{\mathsf{ad}},\\
&\text{(A.2)} \qquad & |F(y_1,u) - F(y_2,u)| \le L_1\,|y_1 - y_2| \qquad & \text{for all } y_1, y_2 \in \mathbb{R}^n, u \in U_{\mathsf{ad}},\\
&\text{(A.3)} \qquad & |L(y_1,u) - L(y_2,u)| \le L_2\,|y_1 - y_2| \qquad & \text{for all } y_1, y_2 \in \mathbb{R}^n, u \in U_{\mathsf{ad}}.
\end{aligned}$$

We note that (A.2) and (A.3) are not required for Proposition A.1. Recall that $\beta = e^{-\lambda h}$ for fixed $\lambda > 0$.

Table 6

Comparisons of results from feedback and open loop design: Discontinuous initial condition.

	Feedback	Open-loop
J w. opt. control	0.0370	0.0353
J w/o control	0.1258	0.1258
Value function	0.0372	

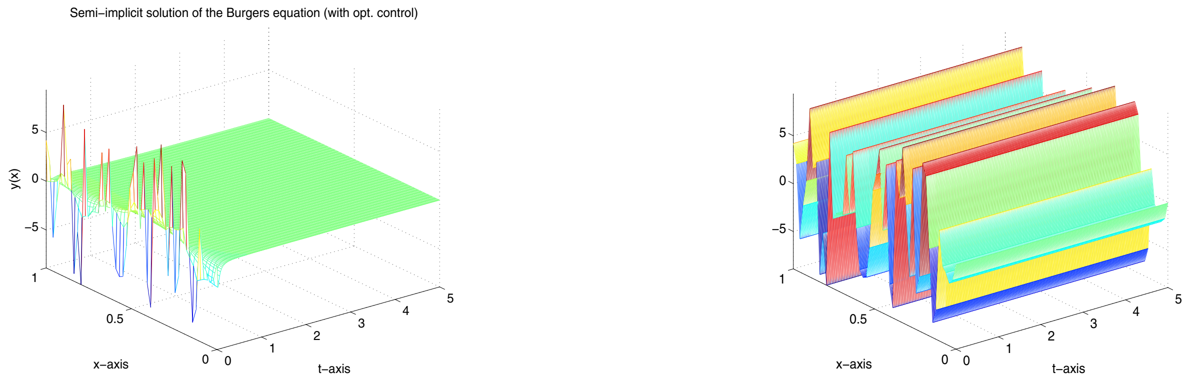


Figure 5. *Optimal state with random noise* (9.0) *in the initial condition: Feedback design (left) and open loop design (right).*

Proposition A.1. *The discrete minimal value function v_h is the unique solution of*

$$v_h(y_\circ) = \inf_{u \in U_{\text{ad}}} \left\{ \frac{h}{2}\big(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\big) + \beta v_h(y_\circ + hF(y_\circ, u)) \right\}. \tag{A.4}$$

Moreover, $|v_h(y_\circ)| \le M(\frac{1}{1-\beta} - \frac{1}{2})$ for all $y_\circ \in \mathbb{R}^n$.

Proof. For $u_h = \{u_0, u_1, \dots\} \in \mathcal{U}^h_{\text{ad}}$, set $\bar{u}_h = \{u_1, u_2, \dots\}$, and denote by $y_h = \{y_j(y_\circ, u_h)\}_{j=1}^{\infty}$ the corresponding solution to (2.9). Then

$$y_{j+1}(y_\circ, u_h) = y_j(y_1, \bar{u}_h) \quad \text{for } j \ge 0,$$

where $y_1 = y_\circ + hF(y_\circ, u_0)$. It follows that

$$J_h(y_\circ, u_h) = \frac{h}{2}\big(L(y_\circ, u_0) + \beta L(y_1, u_0)\big) + \beta J_h(y_1, \bar{u}_h), \tag{A.5}$$

and consequently

$$v_h(y_\circ) \ge \inf_{u \in U_{\text{ad}}} \left\{ \frac{h}{2}\big(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\big) + \beta v_h(y_\circ + hF(y_\circ, u)) \right\}.$$

Conversely, let $u \in U^h_{\text{ad}}$ and $\varepsilon > 0$ be arbitrary. Then there exists $u_h^{\varepsilon} \in \mathcal{U}_{\text{ad}}$ such that

$$v_h(y_\circ + hF(y_\circ, u)) \ge J_h(y_\circ + hF(y_\circ, u), u_h^{\varepsilon}) - \varepsilon.$$

Using (A.5), we have

$$\beta v_h(y_\circ + hF(y_\circ, u)) \ge J(y_\circ, \hat{u}_h^{\varepsilon}) - \frac{h}{2}\big(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\big) - \beta\varepsilon,$$

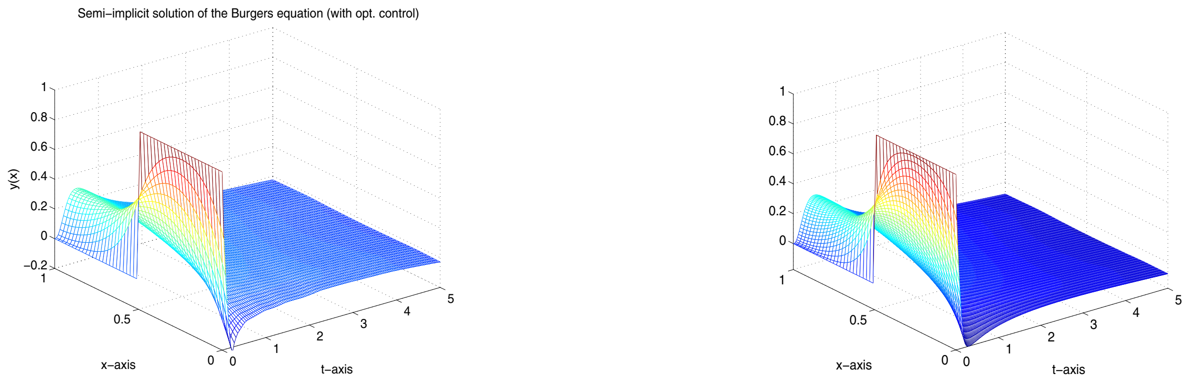


Figure 6. *Optimal state with random noise* (0.25) *in the RHS: Feedback design (left) and open loop design (right).*

where $\hat{u}_h^\varepsilon = \{u, u_0^\varepsilon, u_1^\varepsilon, \dots\}$. This implies that

$$v_h(y_\circ) \le \frac{h}{2}\left(L(y_\circ, u) + \beta L(y_\circ + hF(y_\circ, u), u)\right) + \beta v_h(y_\circ + hF(y_\circ, u)). \tag{A.6}$$

Hence, v_h satisfies (A.4).

Turning to uniqueness, assume that v_h and w_h are two solutions to (A.4) and choose $\varepsilon > 0$ arbitrarily. Then for every $y \in \mathbb{R}^n$ there exists $u^\varepsilon = u^\varepsilon(y) \in U_{\mathsf{ad}}$ such that

$$v_h(y) \ge \frac{h}{2}\left(L(y, u^\varepsilon) + \beta L(y + hF(y, u^\varepsilon), u^\varepsilon)\right) + \beta v_h(y + hF(y, u^\varepsilon)) - \varepsilon,$$

and

$$w_h(y) \le \frac{h}{2}\left(L(y, u^\varepsilon) + \beta L(y + hF(y, u^\varepsilon), u^\varepsilon)\right) + \beta w_h(y + hF(y, u^\varepsilon)).$$

Consequently,

$$\sup_{y\in\mathbb{R}^n}\left(w_h(y) - v_h(y)\right) \le \beta \sup_{y\in\mathbb{R}^n}\left(w_h(y) - u_h(y)\right) + \varepsilon.$$

This estimate also holds with the roles of v_h and w_h exchanged and hence $v_h = w_h$. Moreover by definition of $J_h(y_\circ, u_h)$ and $|\beta| < 1$ we have

$$|v_h(y_\circ)| \le \frac{M}{2}\left(1 + 2\sum_{j=1}^{\infty}\beta^j\right) = M\left(\frac{1}{1-\beta} - \frac{1}{2}\right). \qquad ■$$

Next, continuity of the discrete minimal value functionals is addressed.

Proposition A.2. *For every $h \in (0, 1]$ the minimal value functional is uniformly continuous.*

Proof. Choose ϵ arbitrarily and determine k such that $2M\sum_{j=k+1}^{\infty}\beta^j < \epsilon$. For every $\bar{y} \in \mathbb{R}^n$ there exists $u^\epsilon \in \mathcal{U}_{\mathsf{ad}}^h$ such that $v_h(\bar{y}) \ge J(\bar{y}, u^\epsilon) - \epsilon$. Consequently,

$$v_h(y) - v_h(\bar{y}) \le J(y, u^\epsilon) - J(\bar{y}, u^\epsilon) + \epsilon \quad \text{for every } y \in \mathbb{R}^n,$$

and by (A.1)

$$\begin{aligned} v_h(y) - v_h(\bar y) \le{}& \frac{1}{2}\left(hL(y,u_0^\epsilon) - hL(\bar y,u_0^\epsilon)\right) \\ &+ \frac{h}{2}\sum_{j=1}^{k}\beta^j |L(y_j(y,u^\epsilon),u_{j-1}^\epsilon) - L(y_j(\bar y,u^\epsilon),u_{j-1}^\epsilon)| \\ &+ \frac{h}{2}\sum_{j=1}^{k}\beta^j |L(y_j(y,u^\epsilon),u_{j}^\epsilon) - L(y_j(\bar y,u^\epsilon),u_{j}^\epsilon)| \\ &+ 2M\sum_{j=k+1}^{\infty}\beta^j + \epsilon. \end{aligned}$$

By (A.2) and (A.3), therefore,

$$v_h(y) - v_h(\bar y) \le L_2\,|y-\bar y|\sum_{j=0}^{k}(1+L_1)^j + 2\epsilon.$$

Interchanging the roles of y and $\bar y$ the desired conclusion follows. ■

Proposition A.3. *Every selection of controls*

$$u_j^* \in S_h(y_j^*) = \operatorname*{argmax}_{u\in U_{\mathsf{ad}}}\left\{-\frac{h}{2}(L(y_j^*,u) + \beta L(y_j^* + hF(y_j^*,u),u)) - \beta v_h(y_j^* + hF(y_j^*,u))\right\}$$

with $y_0^* = y_\circ$ and $\{y_j^*\}_{j=1}^\infty$ defined by (2.9) is an optimal feedback control.

Proof. Since U_{ad} is closed and bounded, the mapping $S_h : \mathbb{R}^n \to \mathbb{R}$ is well defined. By (A.5) and the definitions of $\{u_j^*\}_{j=0}^\infty$ and $\{y_j^*\}_{j=0}^\infty$, we have

$$v(y_j^*) = \frac{h}{2}(L(y_j^*,u_j^*) + \beta L(y_{j+1}^*,u_j^*)) + \beta v(y_{j+1}^*)$$

for $j = 0,1,\ldots$. This implies

$$\begin{aligned} &\sum_{j=0}^{\infty}\beta^j(v(y_j^*) - v(y_{j+1}^*)) \\ \text{(A.7)}\qquad &= \frac{h}{2}\sum_{j=0}^{\infty}\beta^j(L(y_j^*,u_j^*) + \beta L(y_{j+1}^*,u_j^*)) \\ &= \frac{h}{2}\left(L(y_\circ,u_0^*) + \sum_{j=1}^{\infty}\beta(L(y_j^*,u_j^*) + L(y_j^*,u_{j-1}^*))\right) = J_h(y_\circ,u_h^*), \end{aligned}$$

and consequently $v(y_\circ) = J_h(y_\circ,u_h^*)$ with $u_h^* = \{u_0^*,u_1^*,\ldots\} \in \mathcal{U}_{\mathsf{ad}}^h$. ■

Proposition A.4. *For every compact set $K \subset U_{\mathsf{ad}}$ we have*

$$\lim_{h\to 0^+}\sup_{y_\circ\in K}|v_h(y_\circ) - v(y_\circ)| = 0,$$

where v is the unique viscosity solution to (HJB).

Proof. The existence of a unique viscosity solution is verified in [4, Theorem III.2.2.], for example. For the convergence result we can proceed as in [4, Theorem VI.1.1.] provided that we verify that $v_h(y_\circ)$ is uniformly bounded w.r.t. $y_\circ$ and $h \in (0, \min(1, 2/\lambda))$; more precisely, we show that

$$\sup\left\{|v_h(y_\circ)| \ : \ y_\circ \in \mathbb{R}^n \text{ and } h \in (0, \min(1,\lambda)]\right\} \le \frac{2M}{\lambda}, \tag{A.8}$$

and that the functions $\underline{v}$ and $\bar{v}$ defined by

$$\underline{v}(y) = \liminf_{(x,h)\to(y,0^+)} v_h(x), \quad \bar{v}(y) = \limsup_{(x,h)\to(y,0^+)} v_h(x) \tag{A.9}$$

are a viscosity supersolution and a viscosity subsolution to (HJB), respectively. To verify (A.8) note that $2M/\lambda$ is a supersolution to (HJB$_h$); i.e., for every $\epsilon > 0$ and $y_\circ \in \mathbb{R}^n$, there exists $u^\epsilon = u^\epsilon(y_o) \in U_{\mathsf{ad}}$ such that

$$\frac{2M}{\lambda} \ge \frac{h}{2}(L(y_\circ, u^\epsilon) + \beta L(y_\circ + hF(y_\circ, u^\epsilon), u^\epsilon)) + \frac{2\beta M}{\lambda} - \epsilon. \tag{A.10}$$

To verify (A.10) we infer from (A.1) and $\beta \le 1$ that

$$\frac{h}{2}(L(y_\circ, u^\epsilon) + \beta L(y_\circ + hF(y_\circ, u^\epsilon), u^\epsilon)) + \frac{2\beta M}{\lambda} \le \frac{M}{\lambda}\,(h\lambda + 2\beta). \tag{A.11}$$

Utilizing $\beta = e^{-\lambda h} \le 1 - \lambda h/2$ for $h \le 2/\lambda$, we find

$$h\lambda + 2\beta \le h\lambda + 2\left(1 - \frac{h\lambda}{2}\right) = 2$$

so that (A.11) implies (A.10). Since v_h is a solution to (HJB), we have

$$v_h(y_\circ) \le \frac{h}{2}(L(y_\circ, u^\epsilon) + \beta L(y_\circ + hF(y_\circ, u^\epsilon), u^\epsilon)) + \beta v_h(y_\circ + F(y_\circ, u^\epsilon)). \tag{A.12}$$

Combining (A.10) and (A.12), we conclude

$$\sup_{y_\circ \in \mathbb{R}^n}\left(v_h(y_\circ) - \frac{2M}{\lambda}\right) \le \beta \sup_{y_\circ \in \mathbb{R}^n}\left(v_h(y_\circ) - \frac{2M}{\lambda}\right) + \epsilon$$

so that $\sup_{y_\circ \in \mathbb{R}^n}(v_h(y_\circ) - 2M/\lambda) \le 0$. Similarly $-2M/\lambda$ is a subsolution of (HJB$_h$). This implies that $\sup_{y_\circ \in \mathbb{R}^n}(-v_h(y_\circ) - 2M/\lambda) \le 0$ and hence (A.8) follows.

To show that $\underline{v}$ is a viscosity supersolution of (HJB), choose $\phi \in C^1(\mathbb{R}^n)$ and let y_1 be a strict minimum of $\underline{v} - \phi$ in the closed ball $\bar{B}(y_1, r), r > 0$. Then (see [4, Lemma V.1.9.]) there exist sequences $\{y_n\}_{n=0}^\infty$ in $\bar{B}(y_1, r)$ and $h_n \to 0^+$ such that

$$(v_{h_n} - \phi)(y_n) = \min_{s \in \bar{B}(y_1, r)} (v_{h_n} - \phi)(s), \ y_n \to y_1, \ v_{h_n}(y_n) \to \underline{v}(y_1). \tag{A.13}$$

Since v_h satisfies (HJB$_h$) we have

$$\begin{aligned}(1-\beta)v_{h_n}(y_n) - \frac{h_n}{2}\left(L(y_n,u_n) + \beta L(y_n + h_n F(y_n,u_n),u_n)\right)\\ +\beta(v_{h_n}(y_n) - \phi(y_n)) - \beta\left(v_{h_n}(y_n + h_n F(y_n,u_n)) - \phi(y_n + h_n F(y_n,u_n))\right)\\ +\beta\left(\phi(y_n) - \phi(y_n + h_n F(y_n,u_n))\right) = 0.\end{aligned}$$

By (A.13) we have for all n sufficiently large that

$$\begin{aligned}(1-\beta)v_{h_n}(y_n) - \frac{h_n}{2}\left(L(y_n,u_n) + \beta L(y_n + h_n F(y_n,u_n),u_n)\right)\\ + \beta\left(\phi(y_n) - \phi(y_n + h_n F(y_n,u_n))\right) \geq 0.\end{aligned}$$

Dividing by h_n and passing to the limit on a subsequence, we obtain

$$\lambda \underline{v}(y_1) - L(y_1,\bar{u}) - \nabla\phi(y_1)\cdot F(y_1,\bar{u}) \geq 0$$

for some $\bar{u} \in U_{\mathsf{ad}}$. Hence $\underline{v}$ is a viscosity supersolution for (HJB). Similarly $\bar{v}$ is a viscosity subsolution. This concludes the proof. ■

REFERENCES

[1] K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, in Shape Optimization and Optimal Design (Cambridge, 1999), Lecture Notes in Pure and Appl. Math. 216, Dekker, New York, 2001, pp. 317–332.

[2] J. A. Atwell, J. T. Borggaard, and B. B. King, *Reduced order controllers for Burgers' equation with a nonlinear observer*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 1311–1330.

[3] H. T. Banks and H. T. Tran, *Reduced order based compensator control of thin film growth in a CVD reactor*, Optimal Control of Complex Structures. Proceedings of the International Conference (Oberwolfach, Germany, 2000), K. H. Hoffmann, et al., eds., *Internat. Ser. Numer. Math.* 139, Birkhäuser, Basel, 2002, pp. 1–17.

[4] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1997.

[5] T. Bewley, P. Moin, and R. Temam, *DNS-based predictive control of turbulence: An optimal benchmark for feedback algorithms*, J. Fluid Mech., 447 (2001), pp. 179–225.

[6] C. I. Byrnes, D. S. Gilliam, and V. I. Shubov, *On the global dynamics of a controlled viscous Burgers' equation*, J. Dynam. Control Systems, 4 (1995), pp. 457–519.

[7] H. Choi, R. Temam, P. Moin, and J. Kim, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[8] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Volume* 5: *Evolution Problems* I, Springer-Verlag, Berlin, 1992.

[9] M. Fahl, E. Arian, and E. W. Sachs, *Trust-Region Proper Orthogonal Decomposition for Flow Control*, NASA/CR-2000-210124, ICASE report 2000-25, ICASE, Hampton, VA, 2000.

[10] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13.

[11] E. Heiberg, *Matlab Parallelization Toolkit*, http://hem.passagen.se/einar_heiberg/.

[12] M. Hinze and S. Volkwein, *Analysis of instantaneous control for the Burgers equation*, Nonlinear Anal., 50 (2002), pp. 1–26.

[13] P. Holmes, J. L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge Monogr. Mech., Cambridge University Press, Cambridge, UK, 1996.

[14] S. Kang, K. Ito, and J. A. Burns, *Unbounded observation and boundary control problems for Burgers equation*, in Proceedings of the 30th IEEE Conference on Decision and Control, IEEE Control Systems Society, Piscataway, NJ, 1991, pp. 2687–2692.

[15] K. Kunisch and S. Volkwein, *Control of Burgers' equation by a reduced-order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1998), pp. 345–371.

[16] F. Leibfritz and S. Volkwein, *Reduced Order Output Feedback Control Design for PDE Systems Using Proper Orthogonal Decomposition and Nonlinear Semidefinite Programming*, Technical report 233, Special Research Center F 003 Optimization and Control, Project area Continuous Optimization and Control, University of Graz & Technical University of Graz, Graz, Austria, submitted.

[17] H. V. Ly, K. D. Mease, and E. S. Titi, *Distributed and boundary control of the viscous Burgers equation*, Numer. Funct. Anal. Optim., 18 (1997), pp. 143–188.

[18] *Message Passing Interface Forum*, http://www.mpi-forum.org/.

[19] J. P. Peraire, W. R. Graham, and K. Y. Tang, *Optimal control of vortex shedding using low order models, Part* 1: *Open-loop model development*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 945–972.

[20] R. D. Prabhu, S. S. Collis, and Y. Cang, *The influence of control on proper orthogonal decomposition of wall-bounded turbulent flows*, Phys. Fluids, 13 (2001), pp. 520–537.

[21] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* I: *Functional Analysis*, Academic Press, New York, 1980.

[22] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.

[23] S. Volkwein, *Optimal control of a phase-field model using the proper orthogonal decomposition*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 83–97.

[24] S. Volkwein, *Lagrange-SQP techniques for the control constrained optimal control problems for the Burgers equation*, Comput. Optim. Appl., 26 (2003), pp. 253–284.