Neural oscillators and integrators

in the dynamics of decision tasks

Eric T. Brown

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Program in

Applied and Computational Mathematics

June 2004

Abstract

In this dissertation I develop both general results on the dynamics of neural oscillators and integrators and specific applications of these results to brain areas involved in simple cognitive tasks. The scientific motivation is broad: neural networks inside our brains are able to adapt to changing information processing demands by exercising cognitive control, for example focussing on salient components of noisy sensory inputs when making specific decisions based on these inputs, but relaxing this focus when other needs become prominent. But what free variables or parameters can account for the observed adaptability? And does this adaptation occur optimally, with respect to simple economic metrics and physiological limitations? Here I address these questions via reduced models of neurons and populations near bifurcations, which characterize the dynamics of a brainstem nucleus involved in adaptive cognitive control, and via variational problems arising from neural signal processing, which clarify the role of this nucleus, and other dynamical mechanisms in decision tasks.

First, I study and apply nonlinear oscillator dynamics. I develop and extend phase reductions for single compartment ordinary differential equation neuron models that show how both type of, and distance from, the four codimension-one bifurcations to periodic firing affect responses of neural populations to stimuli. I also extend results from equivariant dynamics which describe how coupling functions determine the existence and stability of phase-locked states in which subgroups of oscillators are synchronized. These results are then applied to the firing rate dynamics of the locus coeruleus (LC) brainstem nucleus, thereby characterizing the inputs that drive the LC and suggesting a new biophysical mechanism for transitions among LC-mediated states of cognitive performance.

LC-driven neuromodulation transiently adjusts the sensitivity ("gain") of inte-

grator units believed to determine simple cognitive decisions in response to sensory stimuli, and these gain effects are the focus of the second part of this dissertation. I study how transient parameter adjustments can optimize decision tasks for speed and accuracy in the presence of noise. The results indicate a surprising match between empirical data on the time course of LC firing rates and optimal gain trajectories found via variational methods, providing an explicit hypothesis for the role of the LC in decision tasks.

Acknowledgements

No one could ask for a finer advisor than Phil Holmes. I am so grateful for his unceasing encouragement, his guidance in where and how to search for answers, and his high standards for doing and communicating mathematical science. I owe much of my dynamics knowledge directly to his intense and effective teaching. Phil also taught me, in his unique manner that is somehow distinguished, irreverent, passionate, and sincere at once, about the joy that lies at the border of mathematics and science in those wonderful moments when things really come into focus.

I am also grateful to Jonathan Cohen. It has been quite an experience share in Jonathan's inspiration and energy, as well as his insights into how to frame, motivate, visualize, and communicate what I believe are some of the most interesting questions in neuroscience. I feel that Jonathan created an extremely open scientific atmosphere in which I was challenged to achieve my potential, and in which all perspectives and techniques were welcomed and debated. I will miss the pace of our weekly meetings as well as Jonathan's support of me as a person and young scientist.

Jeff Moehlis was like a third advisor to me, and is a wonderful teacher and collaborator. I am grateful to Jeff for his insights, from equivariant dynamics (where he was a great instructor for me) to the practical ups and downs of numerical methods and programming, choosing research problems, and teaching. Jeff also provided an outstanding example for how to live as an effective, happy, and generous young father and (now) professor, which I will in many ways strive to emulate.

I am also grateful to John Hopfield for how seriously he took me every time I came to seek feedback on my work, for challenging me to never be satisfied until I saw the intuitive essence of any neuro-dynamical phenomenon I was trying to study, and for the opportunity to be a student and then TA in his perspective-broadening course. Weinan E always left his door open to me to answer both simple and puzzling questions about stochastic analysis and applied PDEs, and also influenced me through his very useful courses. Yakov Sinai hosted me for hours in his office explaining stochastic processes and provided interesting reading and problems to shore up my knowledge of dynamical systems – I am grateful to him for his generosity and patience in this regard, and also for frequently stunning me with his elegant "two-line" explanations of math phenomena that had me stuck.

It was also a privilege to work with Mark Gilzenrat and Rafal Bogacz, and I am grateful to both for their collaboration, their incisiveness, and their creative energy. I am also grateful for the collaboration with Gary Aston-Jones, Ed Clayton, and Janusz Rajkowski of the Laboratory for Neuromodulation and Behavior at the University of

To my parents

Contents

Chapter 1

Introduction

Recent experimental developments, such as multineuron in vivo recording (e.g. [129, 33]) and functional magnetic resonance imaging (e.g. [28, 105]), are providing high resolution data on neural activity during simple cognitive tasks. One result is an unprecedented set of opportunities to ground theoretical work on the dynamics of neural decisions in the underlying biophysics.

For example, a finding from cortical neural recordings made during behavioral experiments is that certain simple visual discrimination processes can be described by firing rates of neural subpopulations, each of which is responsive to alternative sensory stimuli [18, 83, 156, 72, 145]. These subpopulations appear to accumulate incoming sensory evidence by elevating their firing rates, in addition to receiving both recurrent inputs and excitatory and inhibitory connections from other subpopulations. When the firing rate of the first subpopulation crosses a threshold, a motor response indicating the corresponding decision is made; see Figure 1.1. This makes contact with connectionist theories of cognitive processing, which describe the dynamics of abstracted neural units in cognitive tasks ([148, 134], cf. foundations in [121, 91, 87, 40]), and also with the classical drift-diffusion decision models of Laming, Ratcliff and others (e.g. [113, 139, 140, 159]), since these latter models have recently been shown to be special limits of connectionist networks [169, 21, 13]. Furthermore, Wang [184] has demonstrated how an accumulation of decision-related firing rates can arise from a detailed model of spiking cortical neurons, drawing on prior work on the population dynamics of incoherently firing neural groups (e.g. [165, 63]).

While the cortical recordings just described are striking, they by no means imply that averaged firing rates are the only currency of neural information. In particular, the timing of individual spikes has been shown to play a direct role in encoding certain neural information [12] and has been implemented in novel and powerful computational strategies [92, 93]. This said, I will assume a firing rate description in Part II of this thesis, in which I develop neural integrator models of sensory discrimination tasks of the types for which the experimental data described above was collected. In Part I, I consider biophysically-based neural models that are intrinsic oscillators, and hence preserve spike timing information and allow studies of synchronous oscillations.

A central issue in neuroscience is the control of cognitive processes: the modulation of information processing to enact strategies that accomplish ever-shifting goals
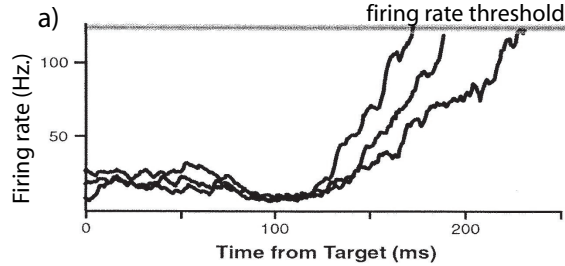
Figure 1.1: Firing rates of area FEF increase to a fixed threshold, at which the decision is made. Note random walk type accumulation of firing rates. From [151].

[137, 153]. A classical example of the need for cognitive control is the tension between 'exploration' and 'exploitation' in which the benefits of flexibly processing a broad variety of sensory and internal signals are balanced against commitment to signals relevant to a particular predetermined task. Recent studies suggest that the forebrain area anterior cingulate cortex (ACC) may be responsible for dynamically assessing the appropriate level of this balance by monitoring the coactivation of neural representations for competing task alternatives and hence the degree of neural *conflict* [27, 15]. High values of this conflict on short timescales indicate the need for additional allocation of cognitive resources to a task in order to, for example, enable more efficient discrimination between alternatives. If this allocation proves unsuccessful, evidenced by maintenance of high conflict levels over longer timescales, in the simplest scenarios the more favorable strategy may be to decrease commitment to the task at hand and explore for more rewarding pursuits.

The locus coeruleus (LC) brainstem nucleus has been proposed as a mechanism for implementing changes in information processing strategy suggested by the time course of conflict (or other means of assessing current attentional needs). Neural recordings show that this nucleus possesses two distinct modes of operation, characterized as *phasic* and *tonic* [10, 167]. In the phasic mode, cortical areas are committed to respond selectively to the limited set of sensory stimuli directly relevant to a given cognitive task: when such stimuli are presented, the LC emits the neuromodulator norepinephrine to widely dispersed brain areas, resulting in a transient increase in the sensitivity (or 'gain' [154]) of cortical units that selectively follows presentation of task-relevant information ([175], cf. [70]). However, baseline levels of norepinephrine release (and hence sensitivity) in the phasic mode are low. The tonic mode, by contrast, is exploratory: there are higher levels of baseline norepinephrine release but smaller stimulus-evoked impulses, allowing responses to a variety of inputs from the environment with less commitment to a specific task.

Several interesting questions remain. For example, how do the changes between tonic and phasic LC modes themselves come about? One possibility, proposed and demonstrated via computations in [167], is that changes in coupling strength between LC neurons result in the different modes. In Chapter 5 of this dissertation I suggest another, simpler mechanism, by which synaptic contacts to LC from the ACC or other brain areas can directly contribute to this change by setting baseline levels

of LC firing. I also address how the varied LC firing patterns observed in different cognitive tasks might arise.

Another key question is: How do we know that transient norepinephrine release can truly be associated with enhancing task-specific processing and throughput of decisions? Answering this question requires modelling cognitive tasks in the presence of dynamic gain manipulations (as in [167, 70]) and assessing the benefits of these manipulations, subjects pursued in Chapters 6 and 7. The results support the hypothesis of [10, 167] that the LC can play a role in optimizing cognitive processing. This hypothesis represents a shift in interpretation of the broader role of the brainstem, which is not typically associated with the details of cognition: in particular, the classical view of norepinephrine is that it sets overall levels of arousal over long timescales, not the timescale of tens and hundreds of milliseconds relevant to cortical decision making.

The dynamics and role of norepinephrine and other neuromodulators in neuroscience is a compelling field of study for additional reasons worth mentioning here. First, such research seeks to identify general principles that may be operative in a wide variety of cognitive tasks. As a consequence, the effects of mechanisms that implement cognitive control must be broadly distributed throughout the brain, implying that a wealth of existing neural data may eventually be applicable and possibly making cognitive control effects easier to identify in future imaging or recording studies. Second, deficits in neuromodulator systems, associated with depleted capacity for, e.g., selective attention and switching between competing task demands, are signatures of most neuropsychiatric diseases [37, 57], adding impetus to the task of understanding the dynamics of these systems and their effects on neural information processing.

Progress on these topics requires understanding the dynamics of the underlying neurons and neural groups on different scales. A first step is to establish generalized (and simplified) neuron models for the dynamics of an isolated spiking neuron, starting with biophysically grounded conductance-based models of Hodgkin-Huxley type [89, 41]. This need for reduced models of this type has been widely recognized and is a broad area of research in mathematical neuroscience. The piecewise-linear 'integrate and fire' formalism, in which voltage increases to a fixed threshold at which a spike is evoked, is perhaps the simplest reduction (e.g. [68]). Another approach taken in developing the Fitzhugh-Nagumo and related relaxation oscillator models [61, 102] is to exploit separations of timescales among gating variables to reduce the dimension of a broad class of conductance based neural models and to develop general approximations that capture the essential components of their dynamics with simple polynomial vectorfields (see also [97]). A third, related approach relies on simplified normal form equations that approximate the full conductance-based equations near bifurcations to periodic firing: this is the 'canonical model' methodology of [94]. A bifurcation-based analysis is also used in, for example, [142] and references therein to explain how broad classes of neuron model can begin firing at either zero or nonzero rates, can undergo bursting behavior, and can be reduced to phase descriptions by projecting their dynamics onto the attracting invariant manifolds on which periodic trajectories lie (see also, e.g., [52, 84, 171, 85, 50, 111]).

3

In Part I of this dissertation I adopt and extend this latter phase reduction approach, thereby contributing to a general analysis of reduced neuron models with applications to the cognitive control applications introduced above. In particular, I address how 'controlling' afferent inputs can adjust the sensitivity of spiking neurons to other stereotyped external (e.g. sensory) inputs, hence allocating different levels of commitment or importance to those stimuli. This is done in Chapter 2 by assessing how the sensitivity of phase models of neural oscillators to transient external inputs scales with their baseline firing rates (which may be determined by the value of the controlling afferents). This requires extensions and additions to the existing literature on phase reductions for the generic codimension-one bifurcations to periodic firing. In Chapter 3, I develop a probabilistic model which explains how the sensitivity of individual phase-reduced neurons scales up to determine the sensitivity of a population of such neurons, and also determine temporal patterns in this population response that are signatures of the different bifurcations. Chapter 4 then focusses on synchrony induced by coupling among members of a neural population rather than external inputs, thereby completing the general description of the dynamics of neural oscillators that is applied to LC neurons, as already mentioned, in Chapter 5.

Part II of this dissertation addresses the impacts that norepinephrine-mediated gain changes (driven by the LC dynamics studied in in Part I) have on the neural integrators whose firing rates are assumed to determine cortical decision processes. Chapter 6 begins by projecting the stochastic network dynamics corresponding to a simple two-alternative choice task onto suitable one-dimensional slow manifolds, and using this reduced representation to solve for the (qualitatively different) gain schedules that optimize decision processing for three commonly used neural integrator models. Next, in Chapter 7, the capacity of the LC to adaptively implement gain schedules that are sufficiently close to these optima in order to improve decision performance is tested numerically by (approximately) solving a series of optimization problems constrained by the neurobiology of the LC. The metric of performance used here is the rate of correct responses (or 'reward rate' [72]) achieved by the simulated decision network. Significant improvements in reward rate are found to accompany LC-mediated gain changes, supporting the hypothesis that the LC plays a role in optimizing the dynamics of neural decisions. I conclude with a brief summary and note some open problems and future directions in Chapter 8. Each chapter is prefaced by a brief summary of its contents and major results.

PART I: Neural Oscillators

Chapter 2

Phase reduction of limit cycle oscillators

2.1 Chapter outline

In this chapter, we compute simplifying reductions to phase variables valid near bifurcations to periodic firing for Hindmarsh-Rose, Hodgkin-Huxley, FitzHugh-Nagumo, and Morris-Lecar single compartment models of neurons, encompassing the four generic (codimension one) bifurcations. Section 2.2 discusses phase reduction techniques for ordinary differential equations with attracting limit cycles. In the following Section 2.3, we recall and compute phase response curves (PRCs) for familiar neuron models near the four codimension-one bifurcations to periodic firing, using normal forms and numerical calculations [51]. Thus, we review part of the broad literature on the topic as well as providing new results: PRCs valid near degenerate Hopf and homoclinic bifurcations, and the scaling of PRCs with the frequency of the neurons from which they are derived. These results will enable much of the analysis in the chapters to follow. Most of the results given here appear in the first part of [23].

2.2 Phase equations for nonlinear oscillators with attracting limit cycles – general considerations

2.2.1 Reduction to phase coordinates

Phase reduction methods have a rich history, including numerous applications in neuroscience. The fundamental coordinate change to phase variables described below originated at least by 1949 [118], with the complementary asymptotic phase ideas expanded in, e.g., [35, 182, 81, 183] and applied in, e.g., [52, 53, 54, 84, 171, 85, 50, 94, 111, 103, 16, 97, 51, 22, 115]; see also the related "spike response method" [69, 68] and references therein.

Our starting point is a general, conductance-based single compartment model of

a neuron:

$$CV\dot{} = \left[I^g(V, \mathbf{n}) + I^b + I(V, t) \right] , \tag{2.1}$$

$$\dot{\mathbf{n}} = \mathbf{N}(V, \mathbf{n}) ; \ (V, \mathbf{n})^T \in \mathbb{R}^d . \tag{2.2}$$

Here V is the voltage difference across the membrane, the $(d-1)$-dimensional vector $\mathbf{n}$ comprises gating variables and $I^g(V, \mathbf{n})$ the associated membrane currents, and C is the cell membrane conductance. The baseline inward current I^b effectively sets oscillator frequency, and will correspond below to a bifurcation parameter. $I(V, t)$ represents synaptic currents from other brain areas due to stimulus presentation; below, we neglect reversal potentials so that $I(V, t) = I(t)$. We write this equation in the general form

$$\dot{\mathbf{x}} = F(\mathbf{x}) + G(\mathbf{x}, t) ; \qquad \mathbf{x} = (V, \mathbf{n})^T \in \mathbb{R}^d , \tag{2.3}$$

where $F(\mathbf{x})$ is the 'baseline' vector field, $G(\mathbf{x}, t)$ is the stimulus effect, and ^{T} denotes transpose. In our simplification, $G(\mathbf{x}, t) = (I(t), \mathbf{0})^T$; in a more general setting, perturbations in the gating equations (2.2) could also be included.

We assume that the baseline ($G \equiv 0$) neural oscillator has a normally hyperbolic [82], attracting limit cycle γ. This limit cycle persists under small perturbations [59], and hereafter we assume that such a limit cycle always exists for each neuron.

The objective is to simplify Eqn. (2.3) by defining a scalar phase variable $\theta(\mathbf{x}) \in [0, 2\pi)$ for all $\mathbf{x}$ in some neighborhood U of γ (within its domain of attraction), such that the phase evolution has the simple form $\frac{d\theta(\mathbf{x})}{dt} = \omega$ for all $\mathbf{x} \in U$ when $G \equiv 0$. Here $\omega = 2\pi/T$, where T is the period of (2.3) with $G \equiv 0$. From the chain rule, this requires

$$\frac{d\theta(\mathbf{x})}{dt} = \frac{\partial\theta}{\partial\mathbf{x}}(\mathbf{x}) \cdot \frac{d\mathbf{x}}{dt}(\mathbf{x}) \bigg|_{G \equiv 0} = \frac{\partial\theta}{\partial\mathbf{x}}(\mathbf{x}) \cdot F(\mathbf{x}) = \omega . \tag{2.4}$$

Eqn. (2.4) defines a first order PDE that the scalar field $\theta(\cdot)$ must satisfy. Here, we describe the 'asymptotic phase' method of constructing $\theta(\cdot)$ *indirectly*, i.e., without solving the PDE itself. We construct this solution in two stages: first, we define $\theta(\mathbf{x})$ for $\mathbf{x} \in \gamma \subset U$, and then we extend to general $\mathbf{x} \in U$.

Let $\mathbf{x}_s$ be the point on the limit cycle γ with the maximum value of V (i.e., the tip of the voltage spike), and let $\theta(\mathbf{x}_s) = 0$. To define θ for the rest of γ, we let $\theta(\mathbf{x}_0^\gamma(t)) = \omega t$, where $\mathbf{x}_0^\gamma(t) \in \gamma$ is the trajectory of the baseline system with $\mathbf{x}_0^\gamma(0) = \mathbf{x}_s$. This gives a simple parameterization of γ by θ. To extend $\theta(\cdot)$ to general $\mathbf{x} \in U$, we define the *isochron* [183] corresponding to a point $\mathbf{x}^\gamma(0) \in \gamma$ to be the set of all initial conditions $\mathbf{x}(0) \in U$ such that the distance between $\mathbf{x}^\gamma(t)$ and $\mathbf{x}(t)$ approaches zero under the vector field F as $t \to \infty$. Such points $\mathbf{x}^\gamma(0)$ and $\mathbf{x}(0)$ are said to have the same *asymptotic phase* $\theta(\mathbf{x}^\gamma(0))$. See Fig. 2.1. Thus, isochrons are level sets of the asymptotic phase. We note that [81], cf. [88, 82], shows that there is a neighborhood of γ foliated by the (unique) family of isochrons, and we restrict our U to lie within this neighborhood.

In fact, isochrons give an *invariant foliation* of U, i.e., the flow map takes isochrons

7
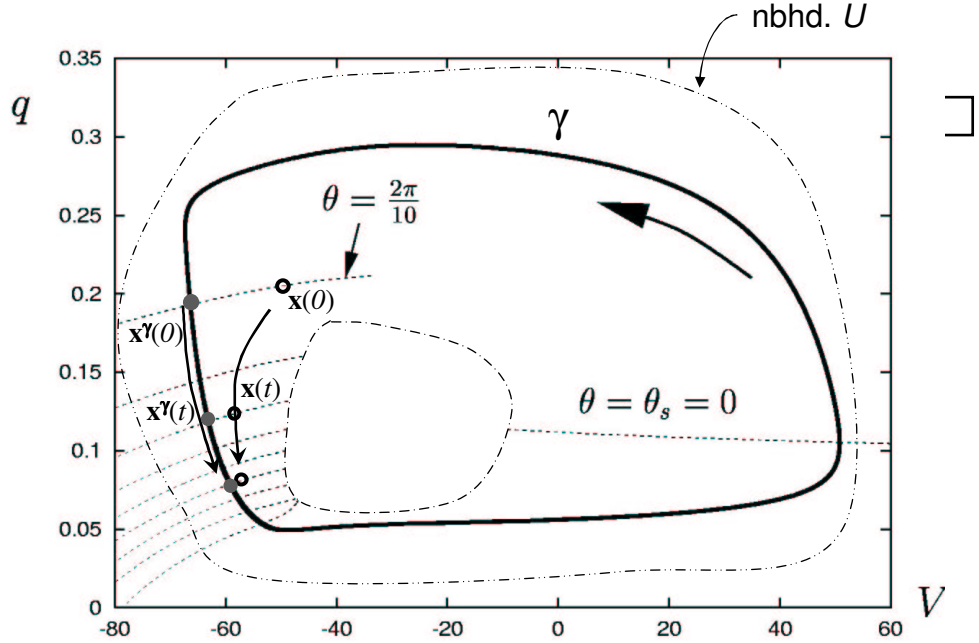
Figure 2.1: The limit cycle γ, annular neighborhood U (between dot-dashed lines), and isochrons (dotted lines) for a planar system. The two trajectories $\mathbf{x}^\gamma(t)$ and $\mathbf{x}(t)$ discussed in Sect. 2.2.1, here with initial asymptotic phase $\theta(\mathbf{x}^\gamma(0)) = \theta(\mathbf{x}(0)) = 2\pi/10$, are shown.

to isochrons. Imagine, for example, two trajectories (as in Fig. 2.1) starting at time $t = 0$ at the points $\mathbf{x}^\gamma(0) \in \gamma$ and $\mathbf{x}(0) \in U$ on the isochron with the asymptotic phase $\theta(\mathbf{x}^\gamma(0)) = \theta(\mathbf{x}(0))$. For all t, the points $\mathbf{x}^\gamma(t)$ and $\mathbf{x}(t)$ must also share an isochron (because the distance between these points also approaches zero under the dynamics as $t \to \infty$), and hence they also have the same asymptotic phase, i.e.,

$$\theta\left(\mathbf{x}^\gamma(t)\right) \equiv \theta\left(\mathbf{x}(t)\right) \quad . \tag{2.5}$$

Differentiating (2.5) with respect to time, and recalling that by definition $\frac{d\theta(\mathbf{x}^\gamma(t))}{dt} = \omega$, we conclude that $\frac{d\theta(\mathbf{x}(t))}{dt} = \omega$ for arbitrary $\mathbf{x}(t)$ in the neighborhood U of γ, as desired.

Finally, we demonstrate that, as in [81], the isochrons must be the level sets of any scalar field $\theta(\cdot)$ in U that solves (2.4). To see this, let L^θ be the level set of such a field with (arbitrary) phase θ, and denote by $\mathbf{x}^\gamma(\theta)$ the intersection of γ and L^θ. L^θ must be mapped into itself after flowing for time T under F, since by assumption $\frac{d\phi}{dt} = \omega = \frac{2\pi}{T}$ (i.e., it is an invariant foliation). Repeating this time-T mapping, we see that all points on L^θ eventually tend to the same point $\mathbf{x}^\gamma(\theta) \in \gamma$ (since L^θ is in the domain of attraction of γ). Thus, all points on L^θ have the same asymptotic phase θ, and L^θ can be none other than an isochron. The solution to (2.4) is therefore determined by the family of isochrons, and is hence (up to an additive constant determined by the choice of origin where $\theta = 0$) unique as claimed.

8

2.2.2 Incorporating inputs $G(\mathbf{x}, t)$

We now re-introduce the stimulus term $G(\mathbf{x}, t)$ and determine its effects on the dynamics of θ. Using the chain rule, as in (2.4) but with $G \neq 0$, we have

$$\frac{d\theta(\mathbf{x})}{dt} = \omega + \frac{\partial \theta}{\partial \mathbf{x}}(\mathbf{x}) \cdot G(\mathbf{x}, t) \ . \tag{2.6}$$

The vector field on right-hand side of (2.6) depends on $\mathbf{x} \in U$, so this equation is not a phase-only (i.e., self-contained) description of the oscillator dynamics. However, evaluating the vector field at $\mathbf{x}^\gamma(\theta)$ (as above, the intersection of γ and the $\theta(\mathbf{x})$ isochron), we have

$$\frac{d\theta(\mathbf{x})}{dt} = \omega + \frac{\partial \theta}{\partial \mathbf{x}}(\mathbf{x}^\gamma(\theta)) \cdot G(\mathbf{x}^\gamma(\theta, t)) + E \ , \tag{2.7}$$

where E is an error term to be discussed below. Dropping this error term, we may rewrite (2.7) as the one-dimensional phase equation

$$\frac{d\theta}{dt} = \omega + \frac{\partial \theta}{\partial \mathbf{x}}(\theta) \cdot G(\theta, t) \ , \tag{2.8}$$

which is valid (up to the error term) in the whole neighborhood U of γ.

Next, we show that the error E in Eqn. (2.7) is of order

$$E = \mathcal{O}\left(|G|^2/\lambda\right) \ ,$$

where $|G|$ measures the magnitude of inputs and λ measures the strength of attraction to the limit cycle γ in a manner to be made precise below. Thus, the phase reduction (2.8) is accurate for sufficiently weak stimuli or sufficiently attracting limit cycles (cf. [183]).

Define $\lambda > 0$ such that $-\lambda$ is an upper bound on exponents characterizing the flow *normal* to γ of the vectorfield F, linearized around any point on γ; thus, λ bounds the (linear) strength of attraction to the limit cycle. In other words, λ is defined such that $v \cdot DF(\mathbf{x}_\gamma)v < -\lambda\|v\|^2$ for any v normal to γ at $\mathbf{x}_\gamma$. (Here and below, $\|\cdot\|$ measures the (Euclidean) distance in $\mathbb{R}^d$.) Also, let the scalar $|G|$ be an upper bound on $G(\mathbf{x}, t)$ over all components as well as over $\mathbf{x}$ and t.

We now show that

$$\|\mathbf{x}(t) - \mathbf{x}_\gamma\| \leq k_1 \cdot |G|/\lambda + O(1/\lambda)^2 \ , \tag{2.9}$$

where $\mathbf{x}_\gamma$ is the intersection of γ and the $\theta(\mathbf{x})$ isochron and k_1 is a positive constant depending on the geometry of the isochrons.

For any $\mathbf{x}$ in the neighborhood of γ, let $\mathbf{x}_b$ be the point on γ closest to $\mathbf{x}$ and define $\mathbf{x}_\perp = \mathbf{x} - \mathbf{x}_b$; see Fig. 2.2. Then we rewrite Eqn. (2.3) as

$$\frac{d\mathbf{x}_b}{dt} + \frac{d\mathbf{x}_\perp}{dt} = F_\parallel(\mathbf{x}_b + \mathbf{x}_\perp) + F_\perp(\mathbf{x}_b + \mathbf{x}_\perp) + G_\parallel(\mathbf{x}_b + \mathbf{x}_\perp) + G_\perp(\mathbf{x}_b + \mathbf{x}_\perp) \ , \tag{2.10}$$

where the vectorfield has been (locally) split into components parallel ($\frac{d\mathbf{x}_b}{dt}$, $F_\parallel$, $G_\parallel$) and orthogonal ($\frac{d\mathbf{x}_\perp}{dt}$, $F_\perp$, $G_\perp$) to γ at $\mathbf{x}_b$. We study the dynamics of the orthogonal component:

$$\frac{d\mathbf{x}_\perp}{dt} = F_\perp(\mathbf{x}_b + \mathbf{x}_\perp) + G_\perp(\mathbf{x}_b + \mathbf{x}_\perp) \ . \tag{2.11}$$

Anticipating small displacements from γ, we linearize the first term around $\mathbf{x}_b$:

$$\frac{d\mathbf{x}_\perp}{dt} = DF_\perp(\mathbf{x}_b)\mathbf{x}_\perp + G_\perp(\mathbf{x}_b + \mathbf{x}_\perp) + \mathcal{O}(\|\mathbf{x}_\perp\|^2) \ ; \tag{2.12}$$

We define the matrix $A(\mathbf{x}_b) = -\frac{1}{\lambda}DF(\mathbf{x}_\gamma)$, which, recalling our definition of λ, expands any vector normal to γ at $\mathbf{x}_b$ by a factor greater than 1. Introducing the small parameter $\epsilon = 1/\lambda$ and the fast time $\tau = t/\epsilon$ we have

$$\frac{d\mathbf{x}_\perp}{d\tau} = -A(\mathbf{x}_b)\mathbf{x}_\perp + \epsilon G_\perp(\mathbf{x}_b + \mathbf{x}_\perp) + \epsilon\mathcal{O}(\|\mathbf{x}_\perp\|^2) \ . \tag{2.13}$$

Next we expand in powers of ϵ:

$$\mathbf{x}_\perp = \mathbf{x}_{\perp,0} + \epsilon\mathbf{x}_{\perp,1} + \epsilon\mathcal{O}(\|\mathbf{x}_\perp\|^2) \ , \tag{2.14}$$

and substitute (2.14) into (2.13) to obtain at order ϵ^0:

$$\frac{d\mathbf{x}_{\perp,0}}{d\tau} = -A(\mathbf{x}_b)\mathbf{x}_{\perp,0} \ . \tag{2.15}$$

Due to the fact that A expands the $\mathbf{x}_{\perp,0}$ by a factor greater than 1 (by definition), $\|\mathbf{x}_{\perp,0}\| < ke^{-\tau} = ke^{-t/\epsilon}$ for some positive k. Thus, $\|\mathbf{x}_{\perp,0}\| = 0$ to all powers in ϵ. Using this result, at order ϵ^1 we have

$$\frac{d\mathbf{x}_{\perp,1}}{d\tau} = -A(\mathbf{x}_b)\mathbf{x}_{\perp,1} + G_\perp(\mathbf{x}_b) \ . \tag{2.16}$$

Using the same property of A, we have

$$\|\mathbf{x}_{\perp,1}(\tau)\| \ = \ \mathcal{O}(|G|). \tag{2.17}$$

From here, we plug our order ϵ^0 and ϵ^1 results into (2.14). After recalling that $\epsilon = 1/\lambda$ and scaling $\|\mathbf{x}_\perp\|$ by k_1 to account for the fact that $\mathbf{x}_b$ and $\mathbf{x}_\gamma$ may not coincide, we have the desired result (2.9).

Next, we Taylor expand the second term on the right-hand side of (2.6) around $\mathbf{x}^\gamma$:

$$\frac{d\theta(\mathbf{x})}{dt} \ = \ \omega + \left[\frac{\partial\theta}{\partial\mathbf{x}}(\mathbf{x}^\gamma) + D\frac{\partial\theta}{\partial\mathbf{x}}(\mathbf{x}^\gamma)(\mathbf{x} - \mathbf{x}^\gamma) + \mathcal{O}\left((\mathbf{x} - \mathbf{x}^\gamma)^2\right)\right] \cdot \tag{2.18}$$

$$\left[G(\mathbf{x}^\gamma, t) + DG(\mathbf{x}^\gamma, t)(\mathbf{x} - \mathbf{x}^\gamma) + \mathcal{O}\left((\mathbf{x} - \mathbf{x}^\gamma)^2\right)\right] \ . \tag{2.19}$$
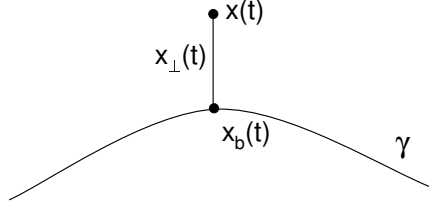
Figure 2.2: Variables used in deriving the bound (2.9).

Comparing with (2.7), we have

$$E = \frac{\partial \theta}{\partial \mathbf{x}}(\mathbf{x}^\gamma) \cdot DG(\mathbf{x}^\gamma, t)(\mathbf{x} - \mathbf{x}^\gamma) + D\frac{\partial \theta}{\partial \mathbf{x}}(\mathbf{x}^\gamma)(\mathbf{x} - \mathbf{x}^\gamma) \cdot G(\mathbf{x}^\gamma, t) + \mathcal{O}\left((\mathbf{x} - \mathbf{x}^\gamma)^2\right) \quad . \quad (2.20)$$

Thus, (2.9) yields $E = \mathcal{O}(|G|^2/\lambda)$ as desired.

2.2.3 Computing the phase response curve

In the case of Eqns. (2.1-2.2), the only partial derivative we must compute to fully define (2.8) is with respect to voltage, and we define the *phase response curve* (PRC) [183] as $\frac{\partial \theta}{\partial V}(\theta) \equiv z(\theta)$. Then, Eqn. (2.8) becomes

$$\frac{d\theta}{dt} = \omega + z(\theta)I(t) \equiv v(\theta, t) \quad , \tag{2.21}$$

the population dynamics of which is the subject of Chapter 3 and is applied to the locus coeruleus in Chapter 5. Note that Eqn. (2.21) neglects reversal potential effects for the various synapses that contribute to the net $I(t)$: if these were included, $I(t)$ would be replaced by $I(\theta, t)$. Furthermore, if G had nonzero components in more than just the voltage direction, we would need to compute a vector-valued PRC; each component of this could be computed in a similar manner to that below.

Direct method

We now describe a straightforward way to compute $z(\theta)$ that is useful in experimental, numerical, and analytical studies. By definition

$$z(\theta) = \lim_{\Delta V \to 0} \frac{\Delta \theta}{\Delta V} \quad , \tag{2.22}$$

where $\Delta \theta = \left[\theta(\mathbf{x}^\gamma + (\Delta V, \mathbf{0})^T) - \theta(\mathbf{x}^\gamma) \right]$ is the change in $\theta(\mathbf{x})$ resulting from a perturbation $V \to V + \Delta V$ from the base point $\mathbf{x}^\gamma$ on γ; see Fig. 2.3. Since $\dot{\theta} = \omega$ everywhere in the neighborhood of γ, the difference $\Delta \theta$ is preserved under the baseline ($G = 0$) phase flow; thus, it may be measured in the limit as $t \to \infty$, when the perturbed trajectory has collapsed back to the limit cycle γ. That is, $z(\theta)$ can

11
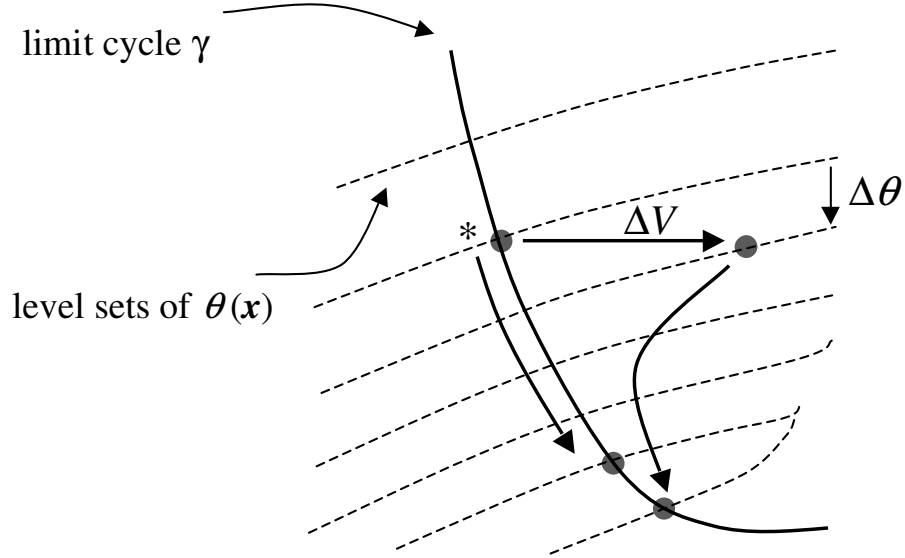
Figure 2.3: The direct method for computing $\frac{\partial \theta}{\partial V}$ at the point indicated by * is to take the limit of $\Delta\theta/\Delta V$ for vanishingly small perturbations ΔV. One can calculate $\Delta\theta$ in the limit $t \to \infty$, as discussed in the text.

be found by comparing the phases of solutions in the infinite-time limit starting on and infinitesimally shifted from base points on γ [183, 71]. This method will be used in Section 2.3 to compute PRCs for the normal forms commonly arising in neural models.

The adjoint method

Another technique for finding $\frac{\partial \theta}{\partial V}(\theta)$ involves solving the adjoint problem associated with Eqns. (2.1-2.2) [94, 54]; this procedure is automated in the program XPP [51] and is equivalent to the direct method discussed above. We note that this equivalence, described below, is implicit in the calculation of coupling functions presented in [94] and [51].

Consider an infinitesimal perturbation $\Delta\mathbf{x}$ to the trajectory $\mathbf{x}^\gamma(t) \in \gamma$ at time $t = 0$. Let $\mathbf{x}(t)$ be the trajectory evolving from this perturbed initial condition. Defining $\Delta\mathbf{x}(t)$ via $\mathbf{x}(t) = \mathbf{x}^\gamma(t) + \Delta\mathbf{x}(t)$,

$$\frac{d\Delta\mathbf{x}(t)}{dt} = DF(\mathbf{x}^\gamma(t))\Delta\mathbf{x}(t) + \mathcal{O}(\|\Delta\mathbf{x}\|^2) \ , \qquad \Delta\mathbf{x}(0) = \Delta\mathbf{x} \ . \qquad (2.23)$$

For the phase shift defined as $\Delta\theta = \theta(\mathbf{x}(t)) - \theta(\mathbf{x}^\gamma(t))$, we have

$$\Delta\theta = \langle \nabla_{\mathbf{x}^\gamma(t)}\theta, \Delta\mathbf{x}(t) \rangle + \mathcal{O}(\|\Delta\mathbf{x}\|^2) \ , \qquad (2.24)$$

where $\langle \cdot, \cdot \rangle$ defines the standard inner product (written as a dot product above),

and $\nabla_{\mathbf{x}^\gamma(t)}\theta$ is the gradient of θ evaluated at $\mathbf{x}^\gamma(t)$. We recall from above that $\Delta\theta$ is independent of time (after the perturbation at $t = 0$) so that taking the time derivative of (2.24) yields, to lowest order in $\|\Delta\mathbf{x}\|$,

$$
\begin{aligned}
\left\langle \frac{d\nabla_{\mathbf{x}^\gamma(t)}\theta}{dt}, \Delta\mathbf{x}(t) \right\rangle &= -\left\langle \nabla_{\mathbf{x}^\gamma(t)}\theta, \frac{d\Delta\mathbf{x}(t)}{dt} \right\rangle \\
&= -\langle \nabla_{\mathbf{x}^\gamma(t)}\theta, DF(\mathbf{x}^\gamma(t))\,\Delta\mathbf{x}(t) \rangle \\
&= -\langle DF^T(\mathbf{x}^\gamma(t))\,\nabla_{\mathbf{x}^\gamma(t)}\theta, \Delta\mathbf{x}(t) \rangle \quad .
\end{aligned}
\tag{2.25}
$$

Here the matrix $DF^T(\mathbf{x}^\gamma(t))$ is the transpose (i.e., adjoint) of the (real) matrix $DF(\mathbf{x}^\gamma(t))$. Since the above equalities hold for *arbitrary* perturbations $\Delta\mathbf{x}(t)$, we have

$$
\frac{d\nabla_{\mathbf{x}^\gamma(t)}\theta}{dt} = -DF^T(\mathbf{x}^\gamma(t))\,\nabla_{\mathbf{x}^\gamma(t)}\theta \quad .
\tag{2.26}
$$

Finally, recall that from (2.4) that

$$
\frac{d\theta}{dt} = \nabla_{\mathbf{x}}\theta \cdot \frac{d\mathbf{x}}{dt} = \nabla_{\mathbf{x}}\theta \cdot F(\mathbf{x}) = \omega \quad ,
\tag{2.27}
$$

which in particular must hold at $t = 0$. Thus, as in [94, 51, 54], we must solve (2.26) subject to the condition

$$
\nabla_{\mathbf{x}^\gamma(0)}\theta \cdot F(\mathbf{x}^\gamma(0)) = \omega \quad .
\tag{2.28}
$$

Since $\nabla_{\mathbf{x}^\gamma(t)}\theta$ evolves in $\mathbb{R}^d$, (2.28) supplies only one of d required initial conditions; the rest arise from requiring that the solution $\nabla_{\mathbf{x}^\gamma(t)}\theta$ to (2.26) be T-periodic [94, 51, 54].

Note that equations (2.26) and (2.28) correspond to equations (9.16) and (9.17) of [94], with the identification of $\nabla_{\mathbf{x}}\theta \to Q$ and a slightly different parametrization. Indeed, this is the adjoint problem that XPP solves to numerically find the PRC Q_{XPP}. The relationship is

$$
\nabla_{\mathbf{x}}\theta = \omega Q_{\mathrm{XPP}} \quad .
\tag{2.29}
$$

The strong attraction method

Since only partial derivatives $\frac{\partial\theta}{\partial\mathbf{x}}$ evaluated on γ enter Eqn. (2.21), and not the value of the phase function θ itself, it is tempting to compute these partial derivatives directly from Eqn. (2.4). However, when viewed as an algebraic equation for the vector field $\frac{\partial\theta}{\partial\mathbf{x}}$, (2.4) yields infinitely many solutions, being only one equation for the d unknown functions $\frac{\partial\theta}{\partial x_j}$, $j = 1, ..., d$. Some of these solutions are much easier to construct than the phase response curve computed via the direct or the adjoint method. However, for such a solution, which we write as $\frac{\partial\theta_2}{\partial\mathbf{x}}$ ($\neq \frac{\partial\theta}{\partial\mathbf{x}}$) to distinguish it from partial derivatives of the asymptotic phase θ, there is *not* necessarily a corresponding phase variable θ_2 such that $\frac{d\theta_2(\mathbf{x})}{dt} = \omega$, $\mathbf{x} \in U$ (in the absence of stimulus): recall the uniqueness of the solution $\theta(\mathbf{x})$ to Eqn. (2.4).

For example, the 'strong attraction limit' of a coordinate change to the phase

variable θ_2 discussed in, e.g., [53, 94] effectively sets

$$\frac{\partial \theta_2}{\partial \mathbf{x}}(\mathbf{x}) = \frac{F(\mathbf{x})}{\|F(\mathbf{x})\|^2}\omega \quad , \tag{2.30}$$

which clearly satisfies (2.4) but implicitly imposes $d-1$ additional constraints: in particular, level sets of θ_2 are always orthogonal to γ, which is not generally the case for isochrons. Furthermore, Eqn. (2.30) requires that $\frac{F(\mathbf{x})}{\|F(\mathbf{x})\|^2}\omega$ is the gradient of the scalar function θ_2, which is only possible if it is curl-free in a neighborhood of γ. Since it is proportional to the unit-normalized vector field which exhibits the attracting limit cycle, $\frac{F(\mathbf{x})}{\|F(\mathbf{x})\|^2}\omega$ will never meet this requirement, so the phase variable θ_2 cannot be extended to a neighborhood of γ. More practically, $\frac{\partial \theta}{\partial \mathbf{x}}(\mathbf{x}^\gamma)$ and $\frac{\partial \theta_2}{\partial \mathbf{x}}(\mathbf{x}^\gamma)$ can also give qualitatively different phase dynamics, with θ dynamics representing more accurately the original 'full' equations: see Chapter 4 for an example involving the stability of phase-locked states in coupled Hodgkin-Huxley systems.

2.2.4 Validity of the phase reduction

We shall always assume that the phase flow $\dot{\theta}$ is nonnegative at the spike point $\theta_s \equiv 0$; otherwise (2.21) does not make sense as a neuron model (neurons cannot cross 'backwards' through the spike and regain a state from which they can immediately fire again). For oscillators giving PRCs $z(\theta)$ with $z(\theta_s) \neq 0$, this assumption restricts admissible perturbing functions $I(t)$ (or, in the more general case of Eqn. (2.8), $G(\mathbf{x}, t)$) to those satisfying

$$I(t)z(\theta_s) > -\omega \quad . \tag{2.31}$$

Thus, for $z(\theta_s) > 0$, excitatory input ($I(t) > 0$) is always admissible, but there is a lower bound on the strength of inhibitory input for which phase reductions hold. In particular, if $I(t)$ contains a noise component, it must be bounded below; this requires 'trimming' the white (diffusive) or Ornstein-Uhlenbeck noise processes commonly used to model variability in synaptic inputs. These problems do not arise for continuous PRCs having $z(\theta_s) = 0$.

We note that $z(\theta_s) = 0$ approximately holds for the Hodgkin-Huxley (HH) and Hindmarsh-Rose (HR) neurons to be considered below, and indeed holds for any neuron model with a 'fast' vector field surrounding the spike tip $\mathbf{x_s}$ on the limit cycle. In this case, asymptotic phase changes very little in a small neighborhood near $\mathbf{x_s}$, since $\theta = \omega t$ and only a short time is spent in the neighborhood. A small perturbation in the V direction therefore takes trajectories to isochrons with similar values of θ, and so has little effect on asymptotic phase. For the integrate and fire systems investigated below, spikes are not explicitly modeled. While this may be viewed as an artificial omission leading to $z(\theta_s) \neq 0$, the population dynamics of such systems are of interest because they are in rather common use.

2.3 Phase equations for nonlinear oscillators with attracting limit cycles – application to neural models

In this section we derive or recall analytical approximations to PRCs for multidimensional systems with limit cycles that arise in the four (local and global) codimension one bifurcations [82]: these are appropriate to conductance-based models of the form (2.1-2.2). We then give PRCs for one-dimensional (linear) 'integrate-and-fire' models. Of these PRC calculations, results for the homoclinic and degenerate Hopf bifurcation are new, while the results for other models, previously derived as referenced in the text, are summarized and recast to display their frequency dependence and for application to population models in what follows.

2.3.1 Phase response curves near codimension one bifurcations to periodic firing

Bifurcation theory [82] identifies four codimension one bifurcations which can give birth to a stable limit cycle for generic families of vector fields: a SNIPER bifurcation (saddle-node bifurcation of fixed points on a periodic orbit), a supercritical Hopf bifurcation, a saddle-node bifurcation of limit cycles, and a homoclinic bifurcation: see Fig. 2.4. All four bifurcation types have been identified in specific neuron models as a parameter, here the baseline inward current I^b, varies: for example, SNIPER bifurcations are found for 'Type I' neurons [50] like the Connor model and its two-dimensional Hindmarsh-Rose (HR) reduction [146], supercritical Hopf bifurcations may occur for the abstracted FitzHugh-Nagumo (FN) model [102], a saddle-node bifurcation of limit cycles is found for the Hodgkin-Huxley (HH) model [89, 143], and a homoclinic bifurcation can occur for the Morris-Lecar (ML) model [142].

In this section, we calculate or summarize PRCs for limit cycles arising from all four bifurcations. This is accomplished, where possible, through use of one- and two-dimensional normal form equations. Normal forms are obtained through center manifold reduction of Eqns. (2.1-2.2) at the bifurcation, followed by a similarity transformation to put the linear part of the equation into Jordan normal form, and finally by successive 'near identity' nonlinear coordinate transformations to remove as many terms as possible, a process which preserves the qualitative dynamics of the system [82]. To obtain the PRC in terms of the original variables, i.e., $\frac{\partial \theta}{\partial V}$, rather than in terms of the normal form variables (which we henceforth denote (x, y)) with associated PRCs $\frac{\partial \theta}{\partial x}$ and $\frac{\partial \theta}{\partial y}$, it is necessary to 'undo' these coordinate transformations. However, since the normal form coordinate transformations only affect nonlinear terms, we obtain the simple relationship

$$\frac{\partial \theta}{\partial V} = \nu_x \frac{\partial \theta}{\partial x} + \nu_y \frac{\partial \theta}{\partial y} + \mathcal{O}(x, y), \tag{2.32}$$

where

$$\nu_x = \left.\frac{\partial x}{\partial V}\right|_{x=y=0}, \qquad \nu_y = \left.\frac{\partial y}{\partial V}\right|_{x=y=0}.$$

The remainder term in (2.32) is assumed to be small near the bifurcations of relevance and is neglected below. This introduces vanishing error in the Hopf case, in which the bifurcating periodic orbits have arbitrarily small radii; the same is true near SNIPER and homoclinic bifurcations, where periodic orbits spend arbitrarily large fractions of their period near the origin. When using the Bautin normal form, however, we must tacitly assume that the nonzero 'onset' radius of stable bifurcating orbits is small; failure of this assumption for the Hodgkin-Huxley model may contribute to the discrepancy between PRCs derived via analytical and numerical methods; see Sect. 2.3.3.

Before proceeding, a few notes regarding the normal form equations that we will consider are in order. For the SNIPER bifurcation, we consider the normal form for a saddle-node bifurcation of fixed points, which must be properly embedded globally in order to capture the presence of the periodic orbit (the unstable branch of the center manifold must close up and limit on the saddle node, cf. Fig. 2.4(a)). For the saddle-node bifurcation of periodic orbits, we appeal to the sequence of bifurcations for 'Type II' neurons such as the Hodgkin-Huxley (HH) model [89], namely a subcritical Hopf bifurcation in which an unstable periodic orbit branch bifurcates from the rest state, turns around, and gains stability in a saddle-node bifurcation of periodic orbits [143]. This sequence is captured by the normal form of the Bautin (degenerate Hopf) bifurcation [112], cf. [82, §7.1]. Finally, for the homoclinic bifurcation we consider only the linearized flow near the fixed point involved in the bifurcation; this is not strictly a normal form, and as for the SNIPER bifurcation, a proper global return interpretation is necessary to produce the periodic orbit.

Near the SNIPER, Hopf, and Bautin local bifurcations, there is a separation of timescales between dynamics along versus dynamics normal to the one-or-two dimensional attracting center manifold containing (or, in the SNIPER case, consisting of) the periodic orbit. In particular, sufficiently close to the bifurcation point, the time required for perturbed solutions to collapse back onto the manifold is negligible compared with the period of the orbit. This implies that, as the bifurcation is approached, (the tangent space of) any $d-1$ dimensional isochron (computed at its intersection with the periodic orbit) becomes normal to the (corresponding tangent space of the) center manifold. Thus, sufficiently near these three bifurcations the only relevant contributions that perturbations make to asymptotic trajectories is via their components *along* the center manifold, as captured by the above terms ν_x and (additionally for the Hopf and Bautin bifurcations) ν_y. Hence Eqn. (2.32) captures the phase response curve for the full d-dimensional system. For the homoclinic *global* bifurcation, the same conclusion holds, although for a different reason: in this case, there is no low dimensional center (i.e. locally slow) manifold. However, because the dynamics which asymptotically determine the PRC are linear for the homoclinic bifurcation (unlike the SNIPER, Hopf, and Bautin cases), a PRC valid for full d-dimensional systems can still be computed analytically, as described below.

We use the direct method of Section 2.2.3 to compute PRCs from the normal form equations. This involves linearizing about the stable periodic orbit, which is appropriate because the perturbations ΔV to be considered are vanishingly small. The explicit solution of the normal form equations yields $\Delta\theta$, and taking limits, we obtain the PRC, cf. (2.22). Without loss of generality, the voltage peak (spike) phase is set at $\theta_s = 0$ and coordinates are defined so that phase increases at a constant rate ω in the absence of external inputs, as in Section 2.2.1. Analogues of some of the following results have been previously derived by alternative methods, as noted in the text, and we also note that PRCs for relaxation oscillators have been discussed in [97]. However, unlike the previous work, here we explicitly compute how the PRCs scale with oscillator frequency.

Saddle-node in a periodic orbit (SNIPER)

A SNIPER bifurcation occurs when a saddle-node bifurcation of fixed points takes place on a periodic orbit: see Fig. 2.4(a). Following the method of [50], we ignore the direction(s) transverse to the periodic orbit, and consider the one-dimensional normal form for a saddle-node bifurcation of fixed points:

$$\dot{x} = \eta + x^2 \ , \tag{2.33}$$

where x may be thought of as local arclength along the periodic orbit. For $\eta > 0$, the solution of (2.33) traverses any interval in finite time; as in [50], the period T of the orbit may be approximated by calculating the total time necessary for the solution to (2.33) to go from $x = -\infty$ to $x = +\infty$ and making the solution periodic by resetting x to $-\infty$ every time it 'fires' at $x = \infty$. This gives $T = \frac{\pi}{\sqrt{\eta}}$, hence $\omega = 2\sqrt{\eta}$.

Since (2.33) is one-dimensional, [50] immediately computes

$$\frac{\partial\theta}{\partial x} = \omega\frac{\partial t}{\partial x} = \frac{\omega}{\frac{dx}{dt}} \ , \tag{2.34}$$

where $\frac{dx}{dt}$ is evaluated on the solution trajectory to (2.33). This gives

$$\frac{\partial\theta}{\partial x} = \frac{2}{\omega}[1 - \cos\theta] \tag{2.35}$$

as first derived in [50], but with explicit ω-dependence displayed here.

Considering a voltage perturbation ΔV, we have

$$\frac{\partial\theta}{\partial V} = z_{SN} = \frac{c_{sn}}{\omega}[1 - \cos\theta] \ , \tag{2.36}$$

where $c_{sn} = 2\nu_x$ is a model-dependent constant (see (2.32) above). Note that $\frac{\partial\theta}{\partial V}$ is nonnegative or nonpositive according to the sign of c_{sn}. Since in 'Type I' neuron models [50] a positive voltage perturbation advances phase (and hence causes the neuron to fire sooner), in the following we will generally assume c_{sn} to be positive.
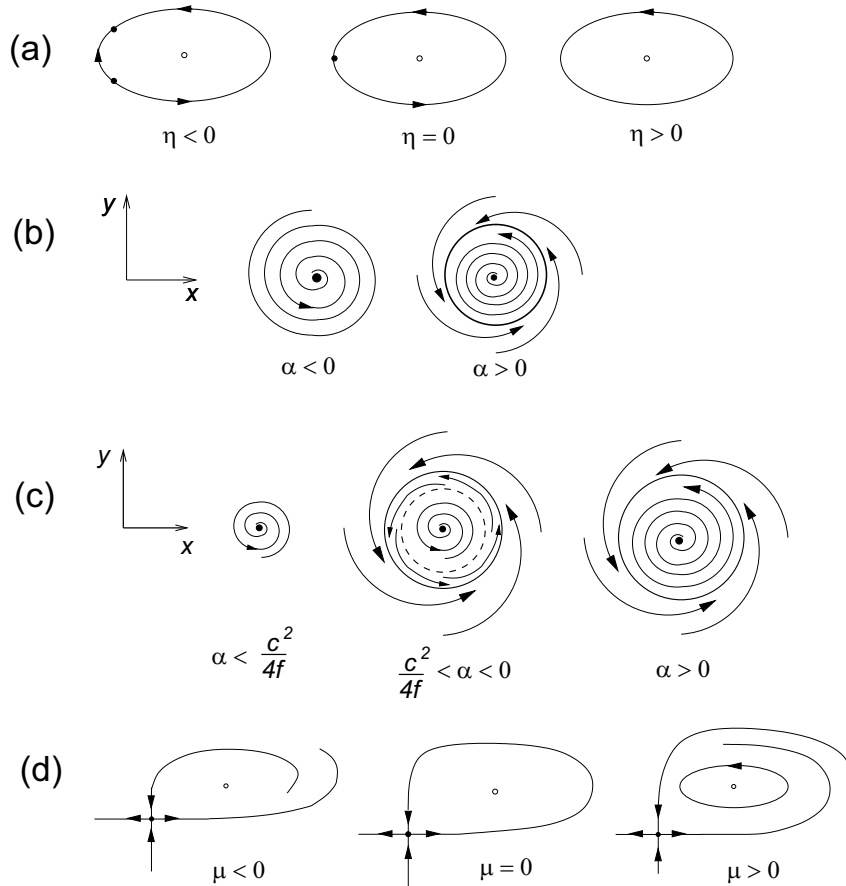
Figure 2.4: (a) SNIPER bifurcation: two fixed points die in a saddle-node bifurcation at $\eta = 0$, giving a periodic orbit for $\eta > 0$, assumed to be stable. (b) Supercritical Hopf bifurcation: a fixed point loses stability as α increases through zero, giving a stable periodic orbit (closed curve). (c) Bautin bifurcation: see text for details. At $\alpha = \frac{c^2}{4f}$ there is a saddle-node bifurcation of periodic orbits. Both a stable (solid closed curve) and unstable (dashed closed curve) periodic orbit exist for $\frac{c^2}{4f} < \alpha < 0$; the unstable periodic orbit dies in a subcritical Hopf bifurcation at $\alpha = 0$. The fixed point is stable (resp., unstable) for $\alpha < 0$ (resp., $\alpha > 0$). (d) Homoclinic bifurcation: a homoclinic orbit exists at $\mu = 0$, giving rise to a stable periodic orbit for $\mu > 0$.

Generalized and supercritical Hopf bifurcations

The normal form for the (generalized) Hopf bifurcation [82, 112] is:

$$\dot{z} = (\alpha + i\beta)z + (c + id)|z|^2 z + (f + ig)|z|^4 z \; ; \tag{2.37}$$

in polar coordinates, this is

$$\dot{r} = \alpha r + c r^3 + f r^5 \; , \tag{2.38}$$
$$\dot{\phi} = \beta + d r^2 + g r^4 \; . \tag{2.39}$$

We study two cases, always treating α as the bifurcation parameter. In the first case, we assume $c < 0$, yielding a supercritical Hopf bifurcation: for $\alpha < 0$ there is a stable fixed point at the origin that loses stability as α increases through zero, giving birth to a stable periodic orbit with radius $r_{po,H} = \sqrt{-\alpha/c}$: see Fig. 2.4(b). Crucially, $r_{po,H} = 0$ when $\alpha = 0$, so that only terms of cubic order in (2.38-2.39) are required to capture (unfold) the supercritical Hopf dynamics. Hence we may set $g = f = 0$ for a local analysis.

In the second case, we assume $c > 0$, so that Eqns. (2.38-2.39) have a subcritical Hopf bifurcation at $\alpha = 0$ and there is no stable periodic orbit for any value of α when $g = f = 0$: hence we must reintroduce these terms to capture the relevant dynamics. Assuming additionally that $f < 0$, for $\alpha < 0$ there is a stable fixed point at the origin that loses stability in a subcritical Hopf bifurcation at $\alpha = 0$, giving rise to an *unstable* periodic orbit as α *decreases* through zero. The branch of unstable periodic orbits turns around at a saddle-node bifurcation of periodic orbits at $\alpha = \frac{c^2}{4f}$; for $\alpha > \frac{c^2}{4f}$ stable periodic solutions exist with radius $r_{po,B} = \left[\frac{1}{2f}\left(-c - \sqrt{c^2 - 4\alpha f}\right)\right]^{1/2}$: see Fig. 2.4(c). This is the generalized Hopf or Bautin bifurcation (identified by the subscript B).

In either case, the angular speed is constant on the stable periodic orbit; hence, we set the asymptotic phase θ equal to the polar angle ϕ on the periodic orbit itself. (However, (radial) level sets of ϕ extending off of the periodic orbit are *not* isochrons, since $\dot{\phi}$ varies with r.)

We calculate the PRC by linearizing about the attracting periodic orbit r_{po}. Letting $r = r_{po} + r'$, we obtain $\dot{r}' = \lambda r' + \mathcal{O}(r'^2)$, where λ is the transverse Floquet exponent (eigenvalue) for the stable periodic orbit. In the supercritical Hopf bifurcation, $\lambda = \lambda_H = -2\alpha < 0$ and $r_{po} = r_{po,H}$; in the Bautin, $\lambda = \lambda_B = \frac{1}{f}\left(c^2 - 4\alpha f + c\sqrt{c^2 - 4\alpha f}\right) < 0$ and $r_{po} = r_{po,B}$. Here and below we drop terms of $\mathcal{O}(r'^2)$ because we are concerned with arbitrarily small perturbations, cf. (2.22). Solving the linearized radial equation with initial condition $r(0) = r_0$, we obtain

$$r(t) = r_{po} + (r_0 - r_{po})e^{\lambda_j t} \; , \tag{2.40}$$

with $j = H$ or B. Next, integrating (2.39) yields

$$\phi(t) = \int_0^t d\phi = \int_0^t [\beta + d(r(s))^2 + g(r(s))^4]ds \ , \tag{2.41}$$

and taking $\phi(0) = \phi_0$, substituting (2.40) in (2.41), letting $t \to \infty$, and dropping terms of $\mathcal{O}(r'^2)$, we obtain the phase θ associated with the initial condition (r_0, ϕ_0):

$$\theta(t) = \phi_0 + \left(\beta + dr_{po}^2 + gr_{po}^4\right) t - \frac{2r_{po}(d + 2gr_{po}^2)(r_0 - r_{po})}{\lambda_B} \ . \tag{2.42}$$

Here we have again used the fact that the polar angle ϕ and the phase θ are identical on the periodic orbit.

Suppose that we start with an initial condition (x_i, y_i) on the periodic orbit, with polar coordinates (r_{po}, ϕ_i). As $t \to \infty$, the trajectory with this initial condition has asymptotic phase $\phi_i + (\beta + dr_{po}^2 + gr_{po}^4)t$. Now consider a perturbation Δx in the x-direction to $(x_f, y_f) = (r_{po} \cos \phi_i + \Delta x, r_{po} \sin \phi_i)$. To lowest order in Δx, this corresponds, in polar coordinates, to

$$(r_f, \phi_f) = \left(r_{po} + \cos \phi_i \Delta x, \ \phi_i - \frac{\sin \phi_i}{r_{po}} \Delta x \right) \ .$$

Setting $(r_0, \phi_0) = (r_f, \phi_f)$ in (2.42) and subtracting the analogous expression with $(r_0, \phi_0) = (r_{po,j}, \phi_i)$, $j = H$ or B, we compute the change in asymptotic phase due to this perturbation:

$$\frac{\partial \theta}{\partial x} = -\frac{2dr_{po,j} + 4gr_{po,j}^3}{\lambda_j} \cos \theta - \frac{1}{r_{po,j}} \sin \theta \ , \tag{2.43}$$

where we have substituted θ for the polar angle ϕ_i, again using the fact that the two variables take identical values on the periodic orbit. Similarly, we find

$$\frac{\partial \theta}{\partial y} = -\frac{2dr_{po,j} + 4gr_{po,j}^3}{\lambda_j} \sin \theta + \frac{1}{r_{po,j}} \cos \theta \ . \tag{2.44}$$

We now express $r_{po,j}$ and λ_j in terms of the frequencies of the periodic orbits. In the supercritical Hopf case (recall that we set $g = f = 0$ here), at the bifurcation point the phase frequency ω is $\dot{\phi} \stackrel{\triangle}{=} \omega_H = \beta$, and from (2.39) we have $\omega - \omega_H = dr_{po,H}^2$, yielding

$$r_{po,H} = \frac{\sqrt{|\omega - \omega_H|}}{\sqrt{|d|}} \ . \tag{2.45}$$

Substituting for $r_{po,H}$, we have $\omega - \omega_H = -\alpha d/c$, which together with the expression for λ_H gives

$$\lambda_H = \frac{2c}{d}(\omega - \omega_H) \ . \tag{2.46}$$

20

In the Bautin case, we find that

$$\omega - \omega_{SN} = \left[-\frac{d}{2f} + \frac{gc}{2f^2} \right] \sqrt{c^2 - 4\alpha f} + \frac{g}{4f^2} \left(c^2 - 4\alpha f \right) , \qquad (2.47)$$

where ω_{SN} is the frequency of the periodic orbit at the saddle-node bifurcation ($\alpha = \frac{c^2}{4f}$). Thus, from (2.47),

$$\sqrt{c^2 - 4\alpha f} = k|\omega - \omega_{SN}| + \mathcal{O}\left(|\omega - \omega_{SN}|^2\right) , \qquad (2.48)$$

where $k = \left| \frac{2f^2}{fd - gc} \right|$, and we may use the expressions for $r_{po,B}$ and λ_B to compute:

$$r_{po,B} = \sqrt{\frac{-c}{2f}} + \mathcal{O}\left(|\omega - \omega_{SN}|\right) , \qquad (2.49)$$

$$\lambda_B = \frac{ck}{f}|\omega - \omega_{SN}| + \mathcal{O}\left(|\omega - \omega_{SN}|^2\right) . \qquad (2.50)$$

Next, we substitute these Eqns. (2.45-2.46) and (2.49-2.50) for r_{po} and λ into (2.43-2.44). For the supercritical Hopf case, this gives

$$\frac{\partial \theta}{\partial x} = \frac{1}{\sqrt{|\omega - \omega_{SN}|}} \frac{\sqrt{|d|}}{|c|} \left[d\cos(\theta) + c\sin(\theta) \right] , \qquad (2.51)$$

$$\frac{\partial \theta}{\partial y} = \frac{1}{\sqrt{|\omega - \omega_{SN}|}} \frac{\sqrt{|d|}}{|c|} \left[d\sin(\theta) - c\cos(\theta) \right] , \qquad (2.52)$$

In the Bautin case, we get

$$\frac{\partial \theta}{\partial x} = \frac{1}{|\omega - \omega_{SN}|} \left[-2d\sqrt{\frac{-c}{2f}} - 4g\left(\frac{-c}{2f}\right)^{3/2} \right] \frac{f}{ck} \cos\theta + \mathcal{O}(1) , \qquad (2.53)$$

$$\frac{\partial \theta}{\partial y} = \frac{1}{|\omega - \omega_{SN}|} \left[-2d\sqrt{\frac{-c}{2f}} - 4g\left(\frac{-c}{2f}\right)^{3/2} \right] \frac{f}{ck} \sin\theta + \mathcal{O}(1) , \qquad (2.54)$$

where we have explicitly written terms of $\mathcal{O}(|\omega - \omega_{SN}|)^{-1}$ which dominate near the saddle-node of periodic orbits. Note that the only term involving the bifurcation parameter α is the prefactor, so that, as this parameter is varied, all other terms in (2.53-2.54) remain constant.

Equipped with (2.51-2.52), the PRC for a perturbation in the V-direction near a supercritical Hopf bifurcation is found from (2.32) to be

$$z_H(\theta) = \frac{\partial \theta}{\partial V} = \frac{c_H}{\sqrt{|\omega - \omega_H|}} \sin(\theta - \phi_H) , \qquad (2.55)$$

where the constant $c_H = \frac{\sqrt{|d|}}{|c|}\sqrt{(\nu_x c + \nu_y d)^2 + (\nu_x d - \nu_y c)^2}$ and the phase shift $\phi_H =$

$\tan^{-1}\left(\frac{\nu_y c - \nu_x d}{\nu_x c + \nu_y d}\right)$. The form of this PRC was originally presented as Eqn. (2.11) of [52]. See that paper, as well as Sect. 4 of [50] and [94], for earlier, alternative methods and computations for the PRC near supercritical Hopf bifurcation.

For the Bautin bifurcation, we similarly arrive at

$$z_B(\theta) = \frac{\partial\theta}{\partial V} = \frac{c_B}{|\omega - \omega_{SN}|} \, \sin(\theta - \phi_B) \, . \tag{2.56}$$

Here $c_B = \left[-2d\sqrt{\frac{-c}{2f}} - 4g\left(\frac{-c}{2f}\right)^{3/2}\right]\frac{f}{ck}\sqrt{\nu_x^2 + \nu_y^2}$ is a constant (which can be positive or negative depending on d and g), and $\phi_B = \tan^{-1}\left(\frac{\nu_x}{\nu_y}\right)$ is an ω-independent phase shift.

Homoclinic bifurcation

Finally, suppose that the neuron model has a parameter μ such that a homoclinic orbit to a hyperbolic saddle point p with real eigenvalues exists at $\mu = 0$. Then there will be a periodic orbit γ for, say, $\mu > 0$, but not for $\mu < 0$. Specifically, we assume a single unstable eigenvalue λ_u smaller in magnitude than that of the all stable eigenvalues, $\lambda_u < |\lambda_{s,j}|$, so that the bifurcating periodic orbit is stable [82]: see Fig. 2.4(d).

If parameters are chosen close to the homoclinic bifurcation, solutions near the periodic orbit spend most of their time near p, where the vector field is dominated by its linearization. This may generically be written in the diagonal form:

$$\dot{x} = \lambda_u x \, , \tag{2.57}$$
$$\dot{y}_j = \lambda_{s,j}\, y_j \, , \; j = 1, ..., d-1 \, , \tag{2.58}$$

where the x and y_j axes are tangent to the unstable and a stable manifold of p, respectively, and $\lambda_{s,j} < 0 < \lambda_u$ are the corresponding eigenvalues. For simplicity, we assume here that the segments of the axes shown in Fig. 2.5 are actually contained in the respective manifolds; this can always be achieved locally by a smooth coordinate change [82].

We define the box $B = [0, \Delta] \times \cdots \times [0, \Delta]$ that encloses γ for the dominant part of its period, but within which (2.57-2.58) is still a good approximation; Δ is model-dependent but fixed for different periodic orbits occurring as a bifurcation parameter varies within the model. We do not explicitly model γ outside of B, but note that the trajectory is 're-injected' after negligible time (compared with that spent in B) at a distance ϵ from the stable manifold, where ϵ varies with the bifurcation parameter μ: see Fig. 2.5. Thus, periodic orbits occurring closer to the bifurcation point correspond to lower values of ϵ and have larger periods.

We approximate the period $T(\epsilon)$ as the time that the x coordinate of γ takes to
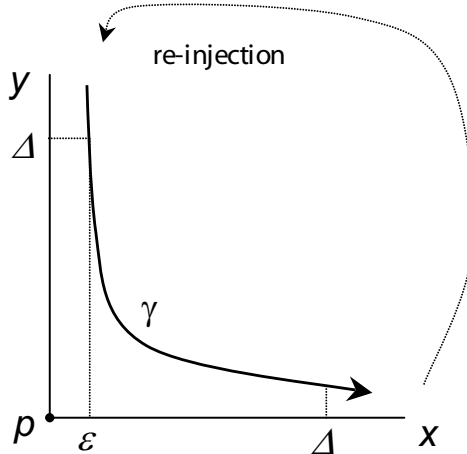
Figure 2.5: The setup for deriving the PRC for oscillations near a homoclinic bifurcation, shown (for simplicity) with $d = 2$.

travel from ϵ to Δ under Eqn. (2.57):

$$T(\epsilon) = \frac{1}{\lambda_u} \ln\left(\frac{\Delta}{\epsilon}\right) . \tag{2.59}$$

Notice that the x-coordinate of γ alone determines $T(\epsilon)$, and hence may be thought of as independently measuring the phase of γ through its cycle. We set $\theta = 0$ at $x = \epsilon$ and, assuming instantaneous re-injection, $\theta = 2\pi$ at $x = \Delta$. Then $\omega = 2\pi/T(\epsilon)$, and as in (2.34)

$$\frac{\partial \theta}{\partial x} = \frac{\omega}{\frac{dx}{dt}} = \frac{\omega}{\lambda_u x(\theta)} = \frac{\omega}{\lambda_u \epsilon} \exp(-\lambda_u \theta/\omega) . \tag{2.60}$$

In the final equality we used the solution to (2.57), $x(t) = \epsilon \exp(\lambda_u t)$, with the substitution $t = \theta/\omega$. Since, as remarked above, motion in the y_j-directions does not affect the phase of γ, only components of a perturbation ΔV along the x-axis contribute to the phase response curve; thus, the PRC $z_{HC} = \frac{\partial \theta}{\partial V} = \nu_x \frac{\partial \theta}{\partial x}$, where ν_x is as defined following (2.32). Using (2.59), $\epsilon = \Delta \exp(-2\pi\lambda_u/\omega)$, which allows us to eliminate ϵ from (2.60):

$$z_{HC}(\theta) = \frac{\partial \theta}{\partial V} = c_{hc}\, \omega \exp\left(\frac{2\pi\lambda_u}{\omega}\right) \exp\left(-\lambda_u \frac{\theta}{\omega}\right) , \tag{2.61}$$

where $c_{hc} = \frac{\nu_x}{\lambda_u \Delta}$ is a model-dependent constant. This is an exponentially decaying

function of θ with maximum

$$z_{max} = c_{hc}\omega \exp\left(\frac{2\pi\lambda_u}{\omega}\right) \tag{2.62}$$

and minimum

$$z_{min} = z_{max} \exp\left(-\frac{2\pi\lambda_u}{\omega}\right) = c_{hc}\omega \ . \tag{2.63}$$

Here and below we assume $c_{hc} > 0$. z_{HC} is discontinuous at the spike point $\theta_s = 2\pi$, which forces us to take a limit in defining population-averaged firing rates below, but does not otherwise affect the following analysis.

2.3.2 One-dimensional neuron models

Generalized integrate and fire models have the form

$$\dot{V} = F(V) + G(V, t) \ , \tag{2.64}$$

where $V(t)$ is constrained to lie between a reset voltage V_r and a threshold V_{th}, and the following reset dynamics are 'externally' imposed: if $V(t)$ crosses V_{th} from below a spike occurs and $V(t)$ is reset to V_r. Here, nothing is lost in transforming to the single phase equation (2.8); in particular, the error term of (2.7) vanishes. In fact, as noted in, e.g., [48], the crucial quantity $\frac{\partial\theta}{\partial V}$ can be found directly from (2.64) with $G(V, t) \equiv 0$:

$$z(\theta) = \frac{\partial\theta}{\partial V} = \omega\frac{\partial t}{\partial V} = \frac{\omega}{F(V(\theta))} \ , \tag{2.65}$$

where we recall that θ is defined such that $\dot{\theta} = \omega$. In the next two subsections we compute phase response curves for two simple integrate and fire models.

Integrate and fire neuron

We first consider the simplest possible integrate and fire (IF) model:

$$C\dot{V} = (I_b + I(t)) \ ; \ \ V_r = 0 \ , \ V_{th} = 1 \ , \tag{2.66}$$

where I_b is the baseline current, C is membrane capacitance, and $G(V, t) = I(t)$. Hereafter we set $C = 1$ for the IF model. The angular frequency of a baseline ($I(t) = 0$) oscillation is $\omega = 2\pi I_b$, and Eqn. (2.65) gives

$$z_{IF}(\theta) = \frac{\omega}{F(V(\theta))} = \frac{\omega}{I_b} \equiv 2\pi \ . \tag{2.67}$$

Thus, the IF PRC is constant in θ and frequency-independent.

Leaky integrate and fire neuron

Next, we consider the leaky integrate and fire (LIF) model:

$$C\dot{V} = (I_b + g_L(V_L - V) + I(t)) \; ; \quad V_r = 0 \; , \; V_{th} = 1 < V_L + \frac{I_b}{g_L} \; , \tag{2.68}$$

where I_b is the baseline current, $g_L > 0$ and V_L are the leak conductance and reversal potential, C is the capacitance, and $G(V,t) = I(t)$. As above, we also set $C = 1$ for this model. We assume $I_b \geq g_L(1 - V_L)$ so that, when $I(t) = 0$, the neuron fires periodically with frequency

$$\omega = 2\pi g_L \left[\ln\left(\frac{I_b + g_L V_L}{I_b + g_L V_L - g_L}\right)\right]^{-1} \; . \tag{2.69}$$

This expression shows how I_b enters as a bifurcation parameter, with $I_b = g_L(1 - V_L)$ corresponding to the bifurcation point at which $\omega = 0$.

Solving (2.68) for $V(t)$ with initial condition $V(0) = V_r = 0$, and then using $\theta = \omega t$ and Eqn. (2.65), gives

$$z_{LIF}(\theta) = \frac{\omega}{g_L}\left(1 - \exp\left(-\frac{2\pi g_L}{\omega}\right)\right)\exp\left(\frac{g_L\theta}{\omega}\right) \; , \tag{2.70}$$

equivalent to formulas previously derived in [171, 115] and references therein. Thus, the PRC for the LIF model is an exponentially increasing function of θ, with a maximum that decreases with ω:

$$z_{max}(\omega) = \frac{\omega}{g_L}\left(\exp\left(\frac{2\pi g_L}{\omega}\right) - 1\right) \; , \tag{2.71}$$

and minimum

$$z_{min}(\omega) = z_{max}\exp\left(-\frac{2\pi g_L}{\omega}\right) = \frac{\omega}{g_L}(1 - e^{-2\pi g_L/\omega}) \; . \tag{2.72}$$

Recall that the PRC near a homoclinic bifurcation is also an exponential function, but with opposite slope: this is because both the essential dynamics near a homoclinic bifurcation and the LIF dynamics are linear, while the trajectories accelerate following spikes in the homoclinic case and decelerate in the LIF.

This is our final analytical PRC calculation; we summarize the results derived above in Table 2.1 and Figs. 2.6-2.7.

2.3.3 Accuracy of the analytical PRCs

The range of parameters over which the PRCs of the full neuron models are well approximated by the analytical expressions derived above varies from model to model. One overall limitation noted in [96] is that normal form calculations for the Bautin and supercritical Hopf bifurcation ignore the relaxation nature of the dynamics of typical
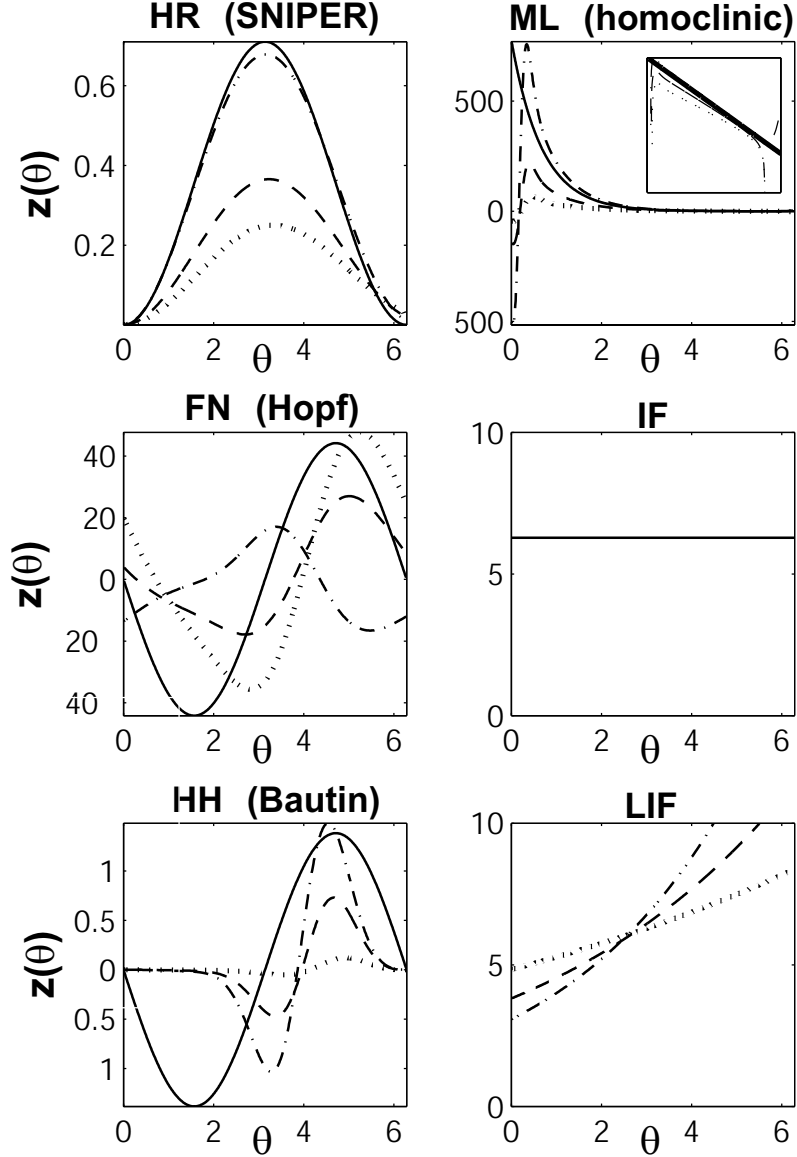
Figure 2.6: PRCs for the various neuron models, from the formulae of Sect. 2.3 and numerically computed using XPP [51], all with $\theta_s = 0$. The relevant bifurcations are noted where applicable. Dot-dashed, dashed and dotted curves for each model correspond to increasing frequencies, respectively: HR: $\omega = 0.0102, 0.0201, 0.0316$ rad/msec (corresp. 1.62, 3.20, 5.03 Hz.) FN: $\omega = 0.204, 0.212, 0.214$ (corresp. 32.5, 33.7, 34.1 Hz.), HH: $\omega = 0.339, 0.355, 0.525$ rad/msec (corresp. 54.2, 56.5, 83.6 Hz.), ML: $\omega = 0.0572, 0.0664, 0.0802$ rad/msec (corresp. 9.10, 10.6, 12.8 Hz.), IF: (any frequency), LIF: $\omega = 0.419, 0.628, 1.26$ rad/msec (corresp. 66.7, 100, 200 Hz.). For the LIF model, $g_L = 0.110$. Normal forms (2.36), (2.55), (2.56), (2.61) for the PRCs closest to bifurcation shown solid (scale factors c_i fit by least-squares); the IF and LIF PRCs are exact. PRC magnitudes decrease with ω for the HR, HH, ML, and LIF models, are constant for the IF model, and increase with ω for the FN model. The phase shifts ϕ_H and ϕ_B are chosen as π (yielding $z(\theta_s) = 0$: see Sect. 2.2.4). The inset to the ML plot displays the same information on a log scale, demonstrating exponential decay.
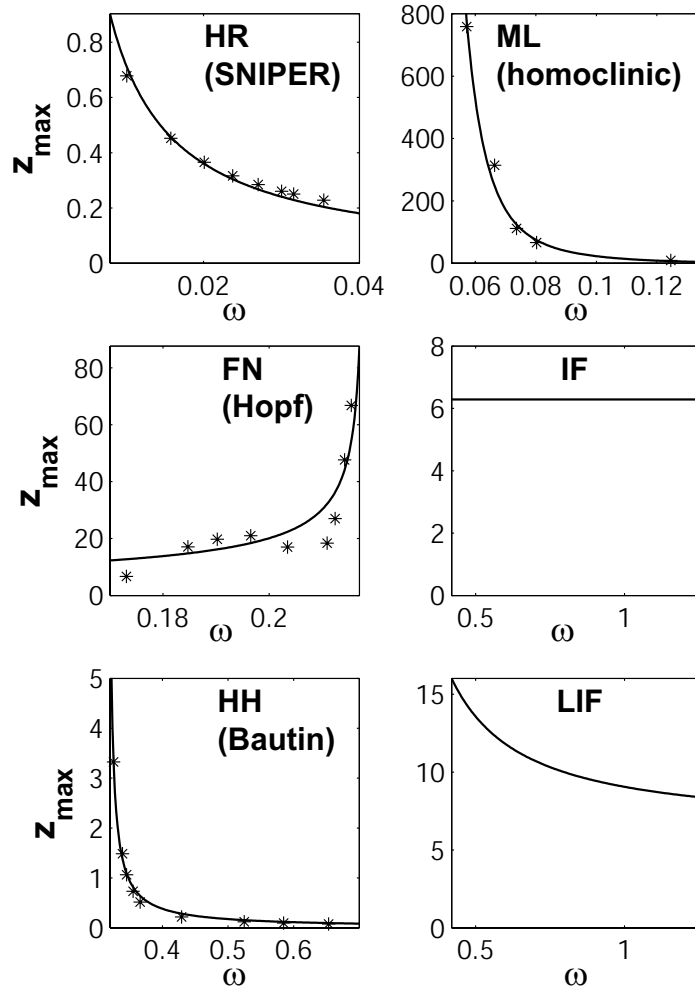
Figure 2.7: Scaling of z_{max} with ω for the various neuron models (and hence scaling of the population response $FL_{max}^d - FL^b$ with ω, see Sect. 3.3.2 below). Asterisks are numerical values from the PRCs of the full HR, FN, HH, and ML models, and curves show predictions of the normal forms (2.36), (2.56), (2.55), (2.61) with least-squares fits (of PRC maxima) for scale factors; results for the IF and LIF models are exact.

bifurcation	$z(\theta)$	z_{max}	z_{min}
SNIPER	$\frac{c_{sn}}{\omega}[1 - \cos(\theta)]$	$\frac{2c_{sn}}{\omega}$	0
Hopf	$\frac{c_H}{\sqrt{|\omega - \omega_H|}}[\sin(\theta - \phi_H)]$	$\frac{c_H}{\sqrt{|\omega - \omega_H|}}$	$-\frac{c_H}{\sqrt{|\omega - \omega_H|}}$
Bautin	$\frac{|c_B|}{|\omega - \omega_{SN}|}[\sin(\theta - \phi_B)] + \mathcal{O}(1)$	$\frac{|c_B|}{|\omega - \omega_{SN}|} + \mathcal{O}(1)$	$-\frac{|c_B|}{|\omega - \omega_{SN}|} + \mathcal{O}(1)$
homoclinic	$c_{hc}\,\omega \exp\left(\frac{2\pi\lambda_u}{\omega}\right) \exp\left(-\lambda_u\theta/\omega\right)$	$c_{hc}\omega \exp\left(\frac{2\pi\lambda_u}{\omega}\right)$	$c_{hc}\omega$
IF	2π	2π	2π
LIF	$\frac{\omega}{g_L}\left(1 - e^{-2\pi g_L/\omega}\right) e^{g_L\theta/\omega}$	$\frac{\omega}{g_L}\left(e^{2\pi g_L/\omega} - 1\right)$	$\frac{\omega}{g_L}(1 - e^{-2\pi g_L/\omega})$

Table 2.1: Phase response curves for the different neuron models.

neural oscillators. However, the analytical PRCs (2.36), (2.55), (2.56), and (2.61) are qualitatively, and in many cases, sufficiently close to the bifurcation point, quantitatively correct: see Fig. 2.6, which compares these formulas with PRCs calculated using XPP [51] for the Hindmarsh-Rose (HR), FitzHugh-Nagumo (FN), Hodgkin-Huxley (HH), and Morris-Lecar (ML) models near the relevant bifurcations (PRCs for the integrate and fire (IF, LIF) models are exact). The companion Fig. 2.7 demonstrates the scaling of PRC maxima with baseline frequency, which is also correctly predicted by the normal form analysis. Frequencies ω were varied by changing the bifurcation parameter: baseline inward current I^b. Here and elsewhere, the neural models are as given in [146, 127, 89], and [142]; all parameter values used here are reproduced along with the equations in the Appendix to Part I. Finally, looking forward to the next chapter, we note that the analytical PRCs derived here will correctly predict key qualitative aspects of population responses to stimuli.

Chapter 3

Response dynamics of phase oscillator populations

3.1 Chapter outline

We undertake a probabilistic analysis of the response of repetitively firing neural populations to simple pulselike stimuli. Equipped with the phase reductions derived in the previous chapter, we use phase density equations to analyze the role of the bifurcation, and the resulting phase response curve (PRC), in response to stimuli. In particular, we explore the interplay between stimulus duration, baseline firing frequency, and population level response patterns. We interpret the results in terms of the signal processing measure of 'gain,' and discuss some further applications and experimentally testable predictions.

This chapter is organized as follows. Following introductory material on population models in neuroscience, Section 3.3 analyzes firing probabilities (and hence population firing rates) in response to simple stimuli, enabling us to predict spike histograms, to describe their dependence on parameters characterizing the stimuli and neuron type, and to emphasize similarities and differences among the responses of different models. These results are summarized in seven Roman-numbered boldface statements. Section 3.4 interprets these results in terms of the gain, or signal amplification, of neural populations, and Section 3.5 analyzes effects of distributed frequencies and noisy currents on the post-stimulus decay of evoked firing rate responses. Section 3.6 closes the chapter with comments on further applications and possible experimental tests. Some of the results of this chapter appear in the second part of [23].

3.2 Introduction and background

This chapter seeks to add to our understanding of how the firing rates of populations of neural oscillators respond to pulselike stimuli representing sensory inputs, and to connect this to mechanisms of neural modulation and computation. In particular, we study how responses depend on oscillator type (classified by its bifurcation to

periodic firing as in Chapter 2), baseline firing rate of the population, and duration of the input. As in, e.g., [60, 86], our results also apply to the interpretation of Peri-Stimulus Time Histograms (PSTHs), which represent averages over an ensemble of independent neuronal recordings.

We are motivated by attempts to understand PSTHs of spike rates in the brainstem organ *locus coeruleus* (see Chapter 5), but there are many other situations in which populations of spiking neurons are reset by stimuli. For example, the multiple-oscillator and beat- frequency models of interval timing of Meck et al. [119] involve cortical oscillators of differing frequencies, and the 40 Hz synchrony reported by Gray and Singer and Eckhorn et al. (see [79, 46] for reviews) also suggest the onset of coherent oscillations in visual cortex.

For most neuron models we find that the response of populations to a fixed stimulus current scales inversely with the pre-stimulus 'baseline' firing rate of the population, reflecting the scaling of PRCs found in Chapter 2. While the firing rates of individual neurons also display this inverse relationship (encoded in their '$f - I$' curves [142]), the scaling of the population response differs from that of individual neurons. This effect suggests a possible role of baseline firing rate in cognitive processing by neural populations: decreasing baseline firing rates (via reduced inputs from other brain areas or via neuromodulators (e.g. [167, 9, 8])) can adjust the 'fraction' of an incoming stimulus that is passed on to the next processing module. Recent data from the brainstem nucleus *locus coeruleus* (LC), for example, reflect this pattern: greater responsivity and better cognitive performance are both correlated with slower baseline firing rates [10, 167, 24], as described in Chapter 5.

We also find that, for certain common neuron models, the maximum population response to a step stimulus of fixed strength can only occur (if it occurs at all) *after* stimulus removal. Moreover, in all cases there are 'resonant' stimulus durations for which there is *no* post-stimulus response. Thus, the magnitude and timing of maximal population response depends strongly both on neuron type *and* stimulus duration relative to baseline period.

Voltage density approaches, primarily undertaken in an integrate and fire framework involving 're-injection' boundary conditions and in some cases involving distributed conductances, are developed and applied in, e.g., [160, 181, 60, 67, 130, 133, 86, 29, 26, 63, 68] and references therein. In particular, density formulations derived from integrate and fire models, e.g. [60, 86], demonstrate the inverse relationship between peak firing rates and baseline frequency (for populations receiving pulsed stimuli) that we extend to other neuron models in this chapter. The work of Brunel et. al [26, 63] focusses on the transmission of stimuli by noisy integrate-and-fire populations: it explains how components of incoming signals are shifted and attenuated (or amplified) when "output" as firing rates of the population, depending on the frequency of the signal component and the characteristics of noise in the population. Some of the conclusions of our chapter (for integrate and fire neurons only) could presumably be reconstructed from the Brunel et al. results by decomposing our stepped stimuli into Fourier components; however, simpler methods applicable to the noise-free case permit different analytical insights into response properties. Experiments on population responses to applied stepped and fluctuating currents have also been

performed, e.g. by [116] in cortical neurons. Due to noise inherent in their biological preparations, responses to stepped, but not fluctuating, stimuli are gradually damped (cf. also [67, 68]); these effects are studied using a phase density approach in [144].

The phase density formulation is also used in [110, 162] and references therein, where the emphasis is on coupling effects in populations with distributed frequencies, generally without external stimuli. The approach closest to ours is that of [163], which focuses on how pulsed input signals can desynchronize populations of noisy, coupled phase oscillators that have clustered equilibrium states; of particular interest is the critical stimulus duration T_{crit} for which the maximum desynchronizing effect is achieved. By contrast, the present chapter focuses on *synchronizing* responses of independent oscillators (with uniform stationary distributions) and, using analytical solutions to this simpler problem in the noiseless limit, stresses the influence of individual neuron properties. Specifically, we contribute a family of simple expressions for time-dependent firing rates in response to pulsed stimuli, derived from different nonlinear oscillator models via phase reductions and the method of characteristics. Our expressions allow us to identify a series of novel relationships between population dynamics during and after stepped stimuli and the frequencies and bifurcation types of the individual neurons making up the population. As already mentioned, we consider only uncoupled neurons, but we note that our results remain generally valid for weakly coupled systems. In particular, in Chapter 5 (and [24]) we show that for a noisy neural population with synaptic and electrotonic couplings sufficient to reproduce observed variations in experimental cross-correlograms, the uncoupled limit is adequate for understanding key 'first order' modulatory effects.

3.3 Probabilistic analysis of firing rates

3.3.1 A phase density equation

We now describe how time-dependent firing rates in response to external stimuli emerge from averages of oscillator population dynamics with appropriate initial conditions. Let $\rho(\theta, t)$ denote the probability density of solutions of (2.21), the fundamental equation describing phase dynamics of an individual neuron from the previous chapter. Thus, $\rho(\theta, t)d\theta$ is the probability that a neuron's phase (say in an arbitrary experimental trial) lies in the interval $[\theta, \theta + d\theta]$ at time t. This density evolves via the advection equation:

$$\frac{\partial \rho(\theta, t)}{\partial t} = -\frac{\partial}{\partial \theta} \left[v(\theta, t) \, \rho(\theta, t) \right] \ . \tag{3.1}$$

Boundary conditions are periodic in the probability flux: e.g., $v(0, t)\rho(0, t) = \lim_{\psi \to 2\pi} v(\psi, t)\rho(\psi, t)$, which reduces to $\rho(0, t) = \rho(2\pi, t)$ for smooth phase response curves z. A related phase density approach is used in [163, 144]. In the presence of noise, there is an additional diffusion term in (3.1) [160, 163, 24]: see Section 3.5 below.

Multiple trials in which stimuli are not keyed to oscillator states may be mod-

eled by averaging solutions of the linear PDE (3.1) over suitably distributed initial conditions; since (unmodeled) noise and variable and/or drifting frequencies tend to distribute phases uniformly in the absence of stimuli, we set $\rho_0 \equiv 1/2\pi$. Histograms of firing times may then be extracted by noting that firing probabilities for arbitrary cells at time t are equal to the passage rate of the probability density through the spike phase, i.e., the probability flux

$$FL(t) \triangleq \lim_{\psi \to \theta_s^-} v(\psi, t)\, \rho(\psi, t) = \lim_{\psi \to \theta_s^-} [\omega + z(\psi)I(t)]\, \rho(\psi, t) \ . \tag{3.2}$$

The limit from below allows for discontinuities in $z(\theta)$ (as in the homoclinic and LIF PRCs of Chapter 2), since the relevant quantity is flux across the spike threshold from lower values of V and hence from lower values of θ. If the PRC $z(\theta)$ and hence $\rho(\theta, t)$ are continuous at θ_s, (3.2) simply becomes $FL(t) = [\omega + z(\theta_s)I(t)]\, \rho(\theta_s, t)$.

We emphasize that the expression (3.2) equally describes the average firing rate of an entire uncoupled population on a single trial, *or* the average firing rate of single neurons drawn from such a population over many sequential trials, as in [86], or a combination of both.

3.3.2 Patterns of firing probabilities and conditions for refractory periods

Eqn. (3.1) can be explicitly solved for piecewise constant stimuli of duration $d = t_2 - t_1$: $I(t) = \bar{I}$ for $t_1 \le t \le t_2$ and $I(t) = 0$ otherwise. (Here and elsewhere we assume $\bar{I} > 0$ unless explicitly noted.) Specifically, the method of characteristics ([176], or pp. 97-100 of [55]) yields:

$$
\begin{aligned}
\rho(\theta, t) &= \rho_0(\Theta_{\theta,t}(0)) \exp\left(-\int_0^t \frac{\partial}{\partial \theta} v(\Theta_{\theta,t}(t'), t')dt' \right) \\
&= \frac{1}{2\pi} \exp\left(-\bar{I} \int_{t_1}^{\tilde{t_2}} z'[\Theta_{\theta,t}(s)]ds \right) \ ,
\end{aligned}
\tag{3.3}
$$

where $t \ge t_1$, $\tilde{t_2} = \min(t, t_2)$ and we take the initial condition $\rho_0 = \rho(\theta, 0) = 1/2\pi$. Here, $\Theta_{\theta,t}(s)$ lies on the characteristic curve given by

$$\frac{d}{ds}\Theta_{\theta,t}(s) = v(\Theta_{\theta,t}(s), s) \ , \tag{3.4}$$

with 'endpoint' condition $\Theta_{\theta,t}(t) = \theta$. When $\Theta_{\theta,t}(s)$ coincides with a discontinuity in z, the integrands in (3.3) are not defined, and we must appeal to the continuity of probability flux or, equivalently, to the following change of variables.

We now simplify the expression (3.3). Using the fact that $v(\Theta_{\theta,t}(s), s) = \omega +$

$\bar{I}z(\Theta_{\theta,t}(s))$ for $t_1 \le s \le t_2$, and changing variables from s to $\Theta_{\theta,t}(s)$,

$$\int_{t_1}^{\tilde{t_2}} z'[\Theta_{\theta,t}(s)]ds = \int_{\Theta_{\theta,t}(t_1)}^{\Theta_{\theta,t}(\tilde{t_2})} \frac{z'[\Theta_{\theta,t}(s)]}{\omega + \bar{I}z(\Theta_{\theta,t}(s))} d\Theta_{\theta,t}(s)$$

$$= \frac{1}{\bar{I}} \ln \left[\frac{\omega + \bar{I}z(\Theta_{\theta,t}(\tilde{t_2}))}{\omega + \bar{I}z(\Theta_{\theta,t}(t_1))} \right] , \tag{3.5}$$

so that

$$\rho(\theta, t) = \frac{1}{2\pi} \left[\frac{\omega + \bar{I}z(\Theta_{\theta,t}(t_1))}{\omega + \bar{I}z(\Theta_{\theta,t}(\tilde{t_2}))} \right] . \tag{3.6}$$

This expression is valid everywhere it is defined. To obtain the terms in (3.6), we integrate (3.4) backward in time from the final condition at $s = t$ until $s = t_1$ or $s = \tilde{t_2}$; this may be done analytically for the normal form PRCs of Sect. 2.3 or numerically for PRCs from full neuron models. The integration yields the PRC-independent expression

$$\Theta_{\theta,t}(\tilde{t_2}) = \theta - \omega(t - \tilde{t_2}) ; \tag{3.7}$$

for all neuron models, while $\Theta_{\theta,t}(t_1)$ is model-dependent via the PRC.

Note that while the stimulus is on (i.e. $t_1 \le t \le t_2$), $\tilde{t_2} = t$ so that $\Theta_{\theta,t}(\tilde{t_2}) = \theta$. After the stimulus turns off, $v(\theta, t)$ is independent of θ, and ρ is constant along curves with constant $\theta - \omega t$. Thus, for $t > t_2$, $\rho(\theta, t)$ is simply a traveling wave rotating with frequency ω, with $\rho(\theta, t_2)$ determining the phase density.

From the definition (3.2), we have:

$$FL(t) = \lim_{\psi \to \theta_s} \frac{\omega + z(\psi)I(t)}{2\pi} \left[\frac{\omega + \bar{I}z(\Theta_{\psi,t}(t_1))}{\omega + \bar{I}z(\Theta_{\psi,t}(\tilde{t_2}))} \right] . \tag{3.8}$$

Fig. 3.1 shows examples of $FL(t)$ for the various neuron models, computed via Eqn. (3.8) with both numerically and analytically derived PRCs z, as well as as via numerical simulations of the full neuron models. The phase reduction (3.8) gives qualitative, and, in some cases, precise matches to the full numerical data. We recall that the accuracy of phase reductions from full neuron models improves with weaker stimuli $\bar{I}$, and that the analytical PRCs better approximate their numerical counterparts as the bifurcation point is approached (i.e., as I_b is varied).

Note that if $\lim_{\psi \to \theta_s} z(\psi) = 0$, $I(t)$ does not directly enter (3.8), so $FL(t)$ depends *only* on variations in ρ *resulting* from the stimulus. However, **(I) if $\lim_{\psi \to \theta_s} z(\psi) \ne 0$, the firing probability $FL(t)$ 'jumps' at stimulus onset and offset;** see Fig. 3.1, and recall that we set $\theta_s = 0$. This is our first main result.

Some comments on the limit in Eqn. (3.8) are appropriate. Since for all neuron models we always assume that $v(\theta)$ is positive and bounded, and is defined except at isolated point(s), $\Theta_{\psi,t}(s)$ is a continuous function of ψ, s and t. Nevertheless, as $\Theta_{\psi,t}(t_1)$ and $\Theta_{\psi,t}(\tilde{t_2})$ pass through θ_s as t advances, discontinuities in $z(\cdot)$ give discontinuities in $FL(t)$, but the limit in Eqn. (3.8) ensures that $FL(t)$ is always defined. As remarked above, if the PRC $z(\cdot)$ is continuous function, then the $\lim_{\psi \to \theta_s} z(\psi) = z(\theta_s)$ and taking the limit is unnecessary.

While the stimulus is on, solutions to (3.4) are periodic with period

$$P = \int_0^{2\pi} \frac{d\theta}{\omega + \bar{I}z(\theta)} \quad , \tag{3.9}$$

(independent of the endpoint condition). Thus, (3.6) implies that $\rho(\theta, t)$ must also be P-periodic, so that the distribution returns to $\rho(\theta, t_1) \equiv \frac{1}{2\pi}$ every P time units: i.e., $\rho(\theta, t_1 + kP) \equiv \frac{1}{2\pi}$ for integers k. If the stimulus is turned off after duration $d = t_2 - t_1 = kP$, this 'flat' density therefore persists (recall that ρ evolves as a traveling wave), giving our second result: **(II) for stimulus durations that are multiples of P, post-stimulus firing probabilities $FL(t)$ return to the constant value $\frac{\omega}{2\pi}$.** This is illustrated in Fig. 3.2 (a) and corresponds to the *absence of post-stimulus refractory periods and ringing*, and is related to the 'black holes' discussed in [163]; Figs. 3.1, 3.2 also illustrate the periodic regimes both during and after the stimulus.

When the stimulus duration d is not a multiple of P (and provided $z(\theta)$ is not constant), $\rho(\theta, t_2)$ has at least one peak exceeding $1/2\pi$, and at least one valley less than $1/2\pi$ (see phase density plots of Fig. 3.1). Let the largest and smallest possible ρ values be ρ_{max} and ρ_{min}, respectively. Eqn. (3.6) then gives

$$\rho_{max} = \frac{1}{2\pi} \left[\frac{\omega + \bar{I}z_{max}}{\omega + \bar{I}z_{min}} \right] \quad ; \quad \rho_{min} = \frac{1}{2\pi} \left[\frac{\omega + \bar{I}z_{min}}{\omega + \bar{I}z_{max}} \right], \tag{3.10}$$

where $z_{min} \equiv z(\theta_{min})$ and $z_{max} \equiv z(\theta_{max})$ are the global extrema of the PRC; note the relationship $\rho_{min}\rho_{max} = 1/4\pi^2$. Recalling that $\Theta_{\theta,t}(\tilde{t}_2) = \theta$ during the stimulus, comparing Eqns. (3.10) and (3.6) shows that ρ_{max} occurs at θ_{min} and ρ_{min} at θ_{max}. When it exists, the stimulus duration d_{max} (resp., d_{min}) for which a distribution with peak ρ_{max} (resp., valley ρ_{min}) occurs is essentially obtained by requiring (ignoring the limits required for discontinuous PRCs) that a characteristic curve passes through θ_{max} (resp., θ_{min}) at t_1 and through θ_{min} (resp., θ_{max}) at time t_2. Thus, **(III) for stimulus durations d_{max} (resp., d_{min}), post-stimulus firing probabilities $FL(t)$ exhibit their maximal deviation above (resp., below) the baseline rate $\frac{\omega}{2\pi}$.** These deviations may or may not be exceeded during the stimulus itself. See Fig. 3.2 for examples and Fig. 3.1 for the evolution of phase density during a prolonged stimulus; in particular, note that while d_{max} is not strictly defined for the LIF model, shorter stimuli (of arbitrarily small duration) always give higher peaks.

We now determine whether maximal peaks and minimal valleys in firing rates occur during or after stimulus for the various neuron types. Again using $\Theta_{\psi,t}(\tilde{t}_2) = \psi$ during the stimulus, (3.8) yields

$$
\begin{aligned}
FL^d(t) &= \lim_{\psi \to \theta_s} \frac{\omega + z(\psi)\bar{I}}{2\pi} \left[\frac{\omega + \bar{I}z(\Theta_{\psi,t}(t_1))}{\omega + \bar{I}z(\psi)} \right] \\
&= \lim_{\psi \to \theta_s} \frac{1}{2\pi} \left[\omega + \bar{I}z(\Theta_{\psi,t}(t_1)) \right] \quad , \quad t_1 < t \le t_2 \quad ;
\end{aligned}
\tag{3.11}
$$

a) HR (SNIPER)

b) FN (Hopf)

c) HH (Bautin)

35
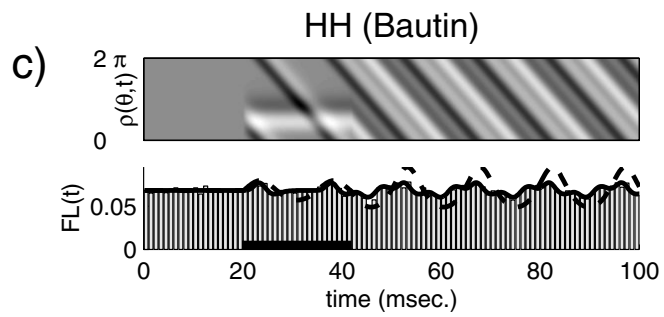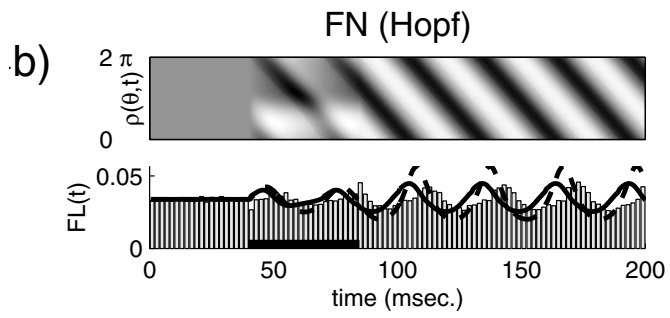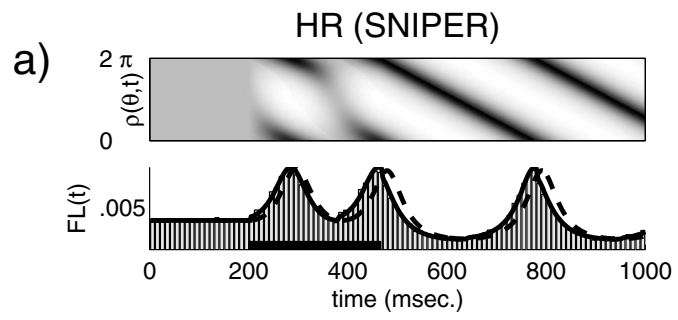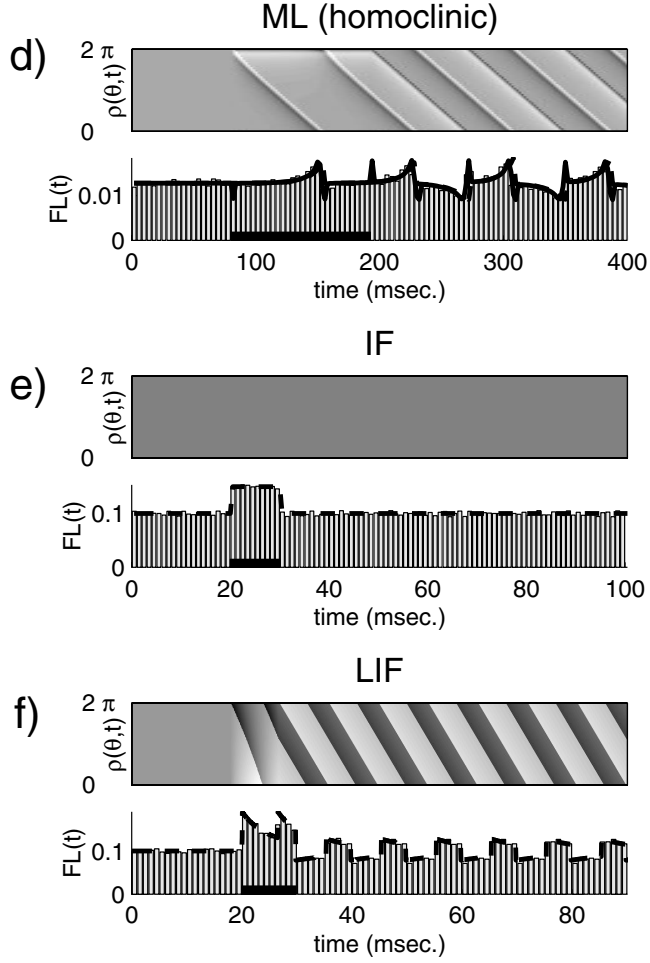
Figure 3.1: (a)-(f) Phase density $\rho(\theta, t)$ in greyscale (darker corresponding to higher values) (top) and firing probability $FL(t)$ in msec^{-1} (bottom) for stimuli of length $3/2 \times P$ (indicated by black horizontal bars), from Eqns. (3.6),(3.8) via the method of characteristics. Dashed curves indicate $FL(t)$ from the normal form PRCs of Eqns. (2.36), (2.55), (2.56), (2.61), (2.67), (2.70); solid curves from numerical PRCs computed via XPP. Baseline frequencies and values of $\bar{I}$ for HR, FN, HH, ML, IF, and LIF models are (0.0201, 0.212, 0.429, 0.08, 0.628, 0.628) rad/msec (corresp. 3.20, 33.7, 68.3, 12.7, 100, 100 Hz.) and (0.1, 0.0015, 0.25, 0.0005, 0.05, 0.05) μA/cm^2, respectively. The vertical bars are PSTHs, numerically computed using the full conductance-based equations (Appendix C) using 10,000 initial conditions, with I_b set to match frequencies of the corresponding phase models. Initial conditions generated by evolving the full equations for a (uniformly distributed) random fraction of their period, from a fixed starting point. Note that $FL(t)$ jumps discontinuously at stimulus onset and offset for the IF and LIF models, since for these models $z(\theta_s) \neq 0$ (point **(I)** in text). Also, during stimulus $FL(t)$ does not dip below the baseline value $\frac{\omega}{2\pi}$ for the HR, IF, and LIF models, because $z_{min} \approx 0$ in these cases (point **V**).

36
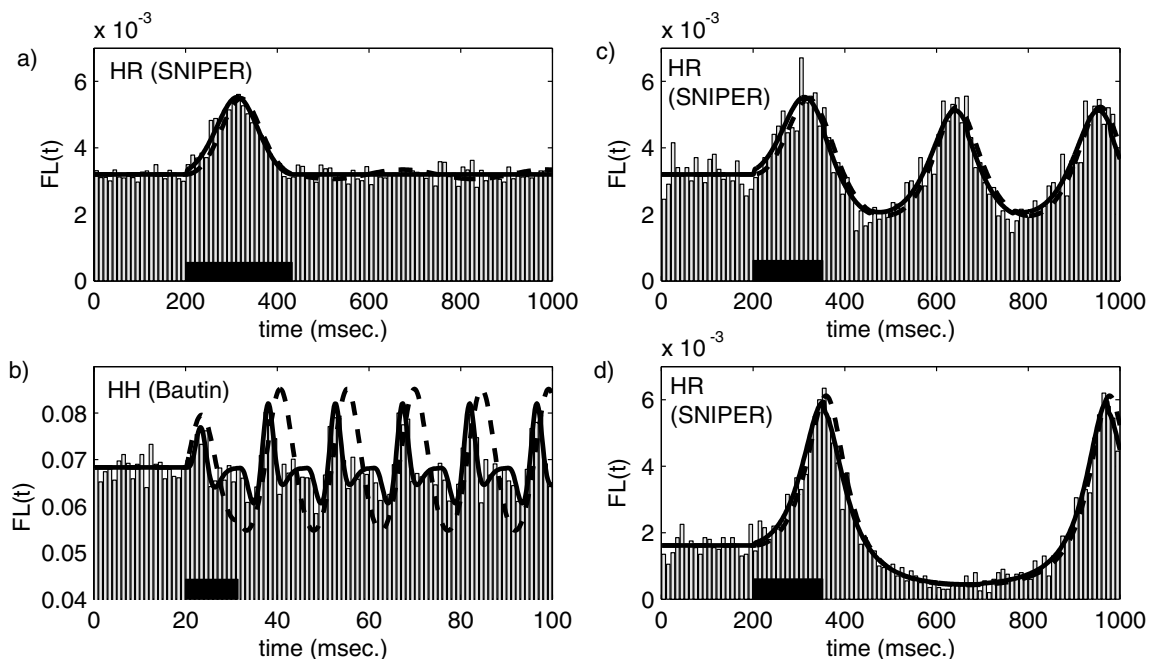
Figure 3.2: (a)-(d) Firing probabilities $FL(t)$ for the HH and HR models, with stimulus characteristics chosen to illustrate the points in the text. Dashed and solid curves and vertical bars denote data obtained as in Fig. 3.1. (a) A stimulus ($\bar{I} = 0.04$ μA/cm^2) of length exactly $P = 232.50$ msec (indicated by the horizontal black bar) for the HR model (ω =0.0201 rad/msec) leaves no trace (point (**II**)). (b) A stimulus ($\bar{I} = 0.25$ μA/cm^2) of duration $d_{max} = 11.46$ msec for the HH model (ω =0.429 rad/msec) yields maximum response *after* the stimulus has switched off (because $z_{min} < 0$) but for the HR model (d) (ω =0.0102 rad/msec) with stimulus duration $d_{max} = 152.01$ msec, the peak in $FL(t)$ is achieved at t_2 (because $z_{min} \approx 0$), (points (**III,IVa**)). Plots (c),(d) illustrate point (**VI**): the stimulus in (c) is identical to that of (d), but the slower HR population (d) (ω=0.0102 vs. 0.0201 rad/msec) displays the greatest response.

the superscript on $FL^d(t)$ denotes 'during' the stimulus, emphasizing that this expression is only valid for $t_1 < t \leq t_2$. After the stimulus has turned off, a different special case of (3.8) is valid:

$$FL^a(t) \;\; = \;\; \lim_{\psi \to \theta_s} \frac{\omega}{2\pi} \left[\frac{\omega + \bar{I}z(\Theta_{\psi,t}(t_1))}{\omega + \bar{I}z(\Theta_{\psi,t}(t_2))} \right] \;\; , \;\; t > t_2 \;\; ; \tag{3.12}$$

here the superscript on $FL^a(t)$ denotes 'after' the stimulus. We now use these expressions to write the maximum and minimum possible firing rates during and after the stimulus:

$$FL^d_{max} \;\; = \;\; \frac{1}{2\pi} \left[\omega + \bar{I}z_{max} \right] \tag{3.13}$$

$$FL^a_{max} \;\; = \;\; \frac{\omega}{2\pi} \left[\frac{\omega + \bar{I}z_{max}}{\omega + \bar{I}z_{min}} \right] \tag{3.14}$$

$$FL^d_{min} \;\; = \;\; \frac{1}{2\pi} \left[\omega + \bar{I}z_{min} \right] \tag{3.15}$$

$$FL^a_{min} \;\; = \;\; \frac{\omega}{2\pi} \left[\frac{\omega + \bar{I}z_{min}}{\omega + \bar{I}z_{max}} \right] \;\; . \tag{3.16}$$

From (3.13-3.16), we have

$$FL^d_{max} - FL^a_{max} \;\; = \;\; \frac{1}{2\pi} \left[\frac{\omega + \bar{I}z_{max}}{\omega + \bar{I}z_{min}} \right] \bar{I}z_{min} \; , \tag{3.17}$$

$$FL^d_{min} - FL^a_{min} \;\; = \;\; \frac{1}{2\pi} \left[\frac{\omega + \bar{I}z_{min}}{\omega + \bar{I}z_{max}} \right] \bar{I}z_{max} \;\; . \tag{3.18}$$

Since we restrict to the case where $v(\theta, t) > 0$ (i.e., there are no fixed points for the phase flow), the terms in the brackets of the preceding equations are always positive. This implies, for $\bar{I} > 0$,

$$FL^a_{max} \;\; \geq \;\; FL^d_{max} \; \text{ if and only if } z_{min} \leq 0 \; , \tag{3.19}$$
$$FL^a_{max} \;\; \leq \;\; FL^d_{max} \; \text{ if and only if } z_{min} \geq 0 \; , \tag{3.20}$$
$$FL^a_{min} \;\; \leq \;\; FL^d_{min} \; \text{ if and only if } z_{max} \geq 0 \; , \tag{3.21}$$
$$FL^a_{min} \;\; \geq \;\; FL^d_{min} \; \text{ if and only if } z_{max} \leq 0 \; , \tag{3.22}$$

where the 'equals' cases of the inequalities require $z_{max} = 0$ or $z_{min} = 0$. In other words, **(IVa) for the specific stimulus durations that elicit maximal peaks in firing rates, these maximal peaks occur during the stimulus if $z_{min} \geq 0$ but after the stimulus switches off if $z_{min} \leq 0$; (IVb) for the specific (possibly different) stimulus durations that elicit minimal firing rate 'dips,' these minimal dips occur during the stimulus if $z_{max} \leq 0$ but after the stimulus switches off if $z_{max} \geq 0$.** We recall that $z_{min} < 0$ is a defining condition for 'Type II' neurons [50]. The post-stimulus maximum (resp. minimum) firing rates are obtained as the peak (resp. valley) of the distribution $\rho(\theta, t)$ passes through θ_s. As Fig. 3.2 (b)

neuron model	Response "jumps" with stimulus? (point **I**)	Max. response *after* stimulus and depressed firing *during* stimulus? (points **IV, V**)
HR	NO	NO
HH	NO	YES
FN	YES	YES
ML	YES	NO
IF	YES	NO
LIF	YES	NO

Table 3.1: Predictions using the numerical PRCs of Fig. 5.3. The conclusions follow from the limiting value of $z(\theta_s)$ (point **(I)** in text), and the value of the PRC minimum z_{min} (points **(IVa, V)**).

shows, the delay from stimulus offset can be significant for typical neuron models.

Defining the baseline rate valid for $t < t_1$

$$FL^b(t) \equiv \frac{\omega}{2\pi} \quad , \tag{3.23}$$

Eqn. (3.15) shows that $FL^d_{min} \geq FL^b$ if and only if $z_{min} \geq 0$. Thus, **(V) if $z_{min} \geq 0$, the firing rate does not dip below baseline values until (possibly) after the stimulus switches off.** Table 3.1 summarizes the above results for the neuron models studied here.

In Section 3.5 we show via Fourier transformation of the analog of Eqn. (3.1) in the presence of noise that $FL(t)$ decays at exponential or faster rates due to noise and averaging over distributions of neuron frequencies (cf. [163, 24]). For *mildly* noisy or heterogeneous systems, the results **(I)-(V)** remain qualitatively similar but are 'smeared:' e.g., $\rho(\theta, t)$ is no longer time-periodic during or after the stimulus, but approaches a generally nonuniform equilibrium state via damped oscillations.

3.3.3 Frequency scaling of response magnitudes

We now determine how the maximum and minimum deviations from baseline firing rates depend on the baseline (pre-stimulus) firing rate of the neural population. Following the discussion of the previous section, we separately compute the scaling of maximal (minimal) responses that are possible *during* stimulus and the scaling of maximal (minimal) responses that are possible *after* stimuli switch off. Eqns. (3.13-

3.16) and (3.23) yield:

$$FL_{max}^d - FL^b = \frac{1}{2\pi}\left[\bar{I} z_{max}\right] \tag{3.24}$$

$$FL_{min}^d - FL^b = \frac{1}{2\pi}\left[\bar{I} z_{min}\right] \tag{3.25}$$

$$FL_{max}^a - FL^b = \frac{\omega}{2\pi}\left[\frac{\bar{I}(z_{max} - z_{min})}{\omega + \bar{I} z_{min}}\right] \tag{3.26}$$

$$FL_{min}^a - FL^b = \frac{\omega}{2\pi}\left[\frac{\bar{I}(z_{min} - z_{max})}{\omega + \bar{I} z_{max}}\right] . \tag{3.27}$$

These expressions provide one set of measures of the sensitivity of population level response at different baseline firing rates. Additionally, taking ratios with the pre-stimulus firing rate (e.g. finding $\frac{FL_{max}^d - FL^b}{FL^b}$) determines the size of deviations relative to baseline activity. We use the information summarized in Table 2.1 to compile these measures for all neuron models in the following Tables 3.2-3.5. Note that in these tables, 'moving away from the bifurcation' means varying parameters so that the frequency varies away from its value at onset of firing, namely $\omega = 0$ for the SNIPER and homoclinic bifurcations and IF and LIF models, ω_H for the supercritical Hopf bifurcation, and ω_{SN} for the Bautin bifurcation. The scaling of $FL_{max}^d - FL^b$, as an example, is confirmed by comparing Fig. 2.7. In summary, **(VI) different neural models and bifurcations imply different scalings of maximal response magnitude with frequency.**

Most measures of population firing rate responses increase for frequencies closer to the bifurcation point (Tables 3.2-3.5). If these models are parameterized so that frequency increases as the bifurcation parameter I_b increases through the bifurcation point, this means that populations at *lower* frequencies tend to display greater responses; see Fig. 2.7 for examples. This effect is further explored in the next section.

3.4 Gain of oscillator populations

In attempts to understand neural information processing, it is useful to understand how input signals are modified by transmission through various populations of spiking cells in different brain organs. The general way to treat this problem is via transfer functions [154, 68]. Here we interpret the results of the previous section in terms of the amplification, or attenuation, of step function input stimuli by the neural population. We consider both extremal and average values of the firing rate $FL(t)$ during stepped stimuli of varying strengths, and illustrate for neurons near a SNIPER bifurcation. We will use the word 'gain' to describe the sensitivity of the resulting input-output relationship: systems with higher gain have a greater output range for a specific set of input strengths. The *average* firing rate during stimulus is

$$\langle FL^d \rangle \equiv \frac{1}{P} , \tag{3.28}$$

bifurcation	$FL^d_{max} - FL^b$	Lowest order scaling near bifurcation	Stronger or weaker effect as move away from bifurcation, to lowest order unnormalized (normalized by FL^b)
SNIPER	$\frac{1}{2\pi}\left[\frac{2\bar{I}c_{sn}}{\omega}\right]$	$\sim \frac{1}{\omega}$	weaker (weaker)
Hopf	$\frac{1}{2\pi}\left[\frac{\bar{I}c_H}{\sqrt{|\omega-\omega_H|}}\right]$	$\sim \frac{1}{\sqrt{|\omega-\omega_H|}}$	weaker (weaker)
Bautin	$\frac{1}{2\pi}\left[\frac{\bar{I}|c_B|}{|\omega-\omega_{SN}|}\right]$	$\sim \frac{1}{|\omega-\omega_{SN}|}$	weaker (weaker)
homoclinic	$\frac{1}{2\pi}\bar{I}c_{hc}\,\omega\exp\left(\frac{2\pi\lambda_u}{\omega}\right)$	$\sim \omega\exp(k/\omega)$	weaker (weaker)
IF	$\bar{I}$	const.	const. (weaker)
LIF	$\frac{1}{2\pi}\frac{\bar{I}\omega}{g_L}\left(e^{2\pi g_L/\omega}-1\right)$	$\sim \omega\exp(k/\omega)$	weaker (weaker)

Table 3.2: Scaling of deviations in firing rate during stimulus $FL^d_{max} - FL^b$ for the different neuron models. The positive constant k differs from case to case.

where P is the period of an individual oscillator during the stimulus (Eqn. (3.9)), and $\langle\cdot\rangle$ is the average over one such period. For the special case of a population near a SNIPER bifurcation, $P_{SN} = \frac{2\pi}{\sqrt{\omega^2+2c_{sn}I}}$ so that

$$\langle FL^d_{SN}\rangle = \frac{\sqrt{\omega^2 + 2c_{sn}I}}{2\pi} \quad . \tag{3.29}$$

These expressions describe the standard '$f - I$' curve typically studied for single neurons [142].

The instantaneous responses of neurons are perhaps of greater interest than averages such as (3.28-3.29). To derive the extremal (i.e., maximally above or below baseline) firing rates, we appeal to the expressions (3.11) and (3.12), which are valid for *both* positive and negative values of $\bar{I}$ as long as $v(\theta,t)$ remains nonnegative. (However, the subsequent formulae of Section 3.3.2 require modification: 'max' and 'min' must be appropriately interchanged when dealing with negative $\bar{I}$.) In particular, the extremal value of $FL^d(t)$ for the SNIPER bifurcation is

$$FL^{d,ext}_{SN} = \frac{1}{2\pi}\left[\omega + \frac{2c_{sn}\bar{I}}{\omega}\right] \quad . \tag{3.30}$$

In Fig. 3.3, we plot $FL^{d,ext}_{SN}$ as a function of both baseline firing rate and stimulus strength $\bar{I}$, where the latter takes both positive and negative values. For (here,

41

bifurcation	$FL^d_{min} - FL^b$	Lowest order scaling near bifurcation	Stronger or weaker effect as move away from bifurcation, to lowest order unnormalized (normalized by FL^b)
SNIPER	0	const.	const. (const.)
Hopf	$-\frac{1}{2\pi}\left[\frac{\bar{I}c_H}{\sqrt{|\omega-\omega_H|}}\right]$	$\sim -\frac{1}{\sqrt{|\omega-\omega_H|}}$	weaker (weaker)
Bautin	$-\frac{1}{2\pi}\left[\frac{\bar{I}|c_B|}{|\omega-\omega_{SN}|}\right]$	$\sim -\frac{1}{|\omega-\omega_{SN}|}$	weaker (weaker)
homoclinic	$\frac{1}{2\pi}\bar{I}c_{hc}\,\omega$	$\sim \omega$	stronger (const.)
IF	$\bar{I}$	const.	const. (weaker)
LIF	$\frac{1}{2\pi}\frac{\bar{I}\omega}{g_L}\left(1-e^{-2\pi g_L/\omega}\right)$	$\sim \omega$	stronger (const.)

Table 3.3: Scaling of deviations in firing rate during stimulus $FL^d_{min} - FL^b$ for the different neuron models.

negative) stimulus values, and frequencies, sufficient to cause the minimum of $v(\theta)$ to dip below zero, fixed points appear in the phase model, giving firing rates $FL^d(t) = \langle FL^d_{SN}\rangle = FL^{d,ext}_{SN} = 0$. Notice the increased sensitivity of extremal firing rates to changes in stimulus strength at low baseline frequencies. This 'increased gain' is also shown in Fig. 3.4 (a), which plots slices through Fig. 3.3 for two different baseline frequencies. However, there is no analogous effect for the average firing rates of Eqn. (3.29), which follow the standard frequency-current relationships for individual neurons: see Fig. 3.4 (b).

Note that there is always a crossing point between firing rate curves for near-SNIPER populations with high and low baseline frequencies (see Fig. 3.4 (a)). Above this crossing point, stimuli are more greatly amplified by the low frequency population; below the crossing point, they are more greatly amplified by the high frequency population. This is analogous to increasing the slope (= gain) of a sigmoidal response function as in [154], gain increase in Fig. 1 of that paper being analogous to decrease of ω. Thus, *if signal discrimination depends on extremal firing rates*, the effects of gain modulation on signal/noise discrimination of [154] could be produced by changes in baseline rate.

bifurcation	$FL^a_{max} - FL^b$	Lowest order scaling near bifurcation	Stronger or weaker effect as move away from bifurcation, to lowest order unnormalized (normalized by FL^b)
SNIPER	$\frac{1}{2\pi}\left[\frac{2\bar{I}c_{sn}}{\omega}\right]$	$\sim \frac{1}{\omega}$	weaker (weaker)
Hopf	$\frac{1}{2\pi}\left[\frac{2\bar{I}c_H\omega}{\omega\sqrt{|\omega-\omega_H|}-\bar{I}c_H}\right]$	$\sim \frac{1}{\sqrt{|\omega-\omega_H|}}$	weaker (weaker)
Bautin	$\frac{1}{2\pi}\left[\frac{2\bar{I}|c_B|\omega}{\omega|\omega-\omega_{SN}|-\bar{I}|c_B|}\right]$	$\sim \frac{1}{|\omega-\omega_{SN}|}$	weaker (weaker)
homoclinic	$\frac{1}{2\pi}\frac{\bar{I}c_{hc}\omega}{1+\bar{I}c_{hc}}\left(\exp(2\pi\lambda_u/\omega)-1\right)$	$\sim \omega\exp(k/\omega)$	weaker (weaker)
IF	0	const.	const. (const.)
LIF	$\frac{\omega}{2\pi}\frac{\bar{I}(1-e^{-2\pi g_L/\omega})(e^{2\pi g_L/\omega}-1)}{g_L+\bar{I}(1-e^{-2\pi g_L/\omega})}$	$\sim \omega\exp(k/\omega)$	weaker (weaker)

Table 3.4: Scaling of deviations in firing rate after stimulus, $FL^a_{max} - FL^b$, for the different neuron models. The positive constant k differs from case to case.

3.5 Effects of noise and distributed frequencies

We now consider two features of realistic neural populations: noise and heterogeneous oscillator frequencies, and determine how these effects damp post-stimulus "ringing" of the population, thereby generalizing formulas in [163] to treat populations with distributed frequencies. Specifically, we assume that baseline frequencies in the neural population (or collection of experimental trials) of interest are distributed with a density $r(\omega)$, so that

$$\langle FL(t)\rangle = \int r(\omega)FL(t,\omega)d\omega \quad . \tag{3.31}$$

Extending Eqns. (2.1)-(2.2) to allow for variable inputs, we consider a population of conductance-based neurons, indexed by i:

$$C\dot{V}_i = \left[I^g(V_i,\mathbf{n}_i)+I^b_i+I(t)+\sigma\eta_i(t)\right] \quad , \tag{3.32}$$
$$\dot{\mathbf{n}}_i = \mathbf{N}(V_i,\mathbf{n}_i) \quad . \tag{3.33}$$

Each neuron may have a different baseline current I^b_i (leading to distributed frequencies across the population) and receives an independent white noise process $\sigma\eta_i(t)$ with r.m.s. strength σ^2 which represents unmodeled 'fast' synaptic inputs. Reducing

bifurcation	$FL^a_{min} - FL^b$	Lowest order scaling near bifurcation	Stronger or weaker effect as move away from bifurcation, to lowest order unnormalized (normalized by FL^b)
SNIPER	$-\frac{1}{2\pi}\left[\frac{2\bar{I}c_{sn}}{\omega+2c_{sn}\bar{I}/\omega}\right]$	$\sim -\omega$	stronger (const.)
Hopf	$-\frac{1}{2\pi}\left[\frac{2\bar{I}c_H\omega}{\omega\sqrt{|\omega-\omega_H|}+\bar{I}c_H}\right]$	$\sim -\frac{1}{\sqrt{|\omega-\omega_H|}}$	weaker (weaker)
Bautin	$-\frac{1}{2\pi}\left[\frac{2\bar{I}|c_B|\omega}{\omega|\omega-\omega_{SN}|+\bar{I}|c_B|}\right]$	$\sim -\frac{1}{|\omega-\omega_{SN}|}$	weaker (weaker)
homoclinic	$\frac{\bar{I}c_{hc}\omega}{2\pi}\frac{\exp\left(-\frac{2\pi\lambda_u}{\omega}\right)-1}{\exp\left(\frac{-2\pi\lambda_u}{\omega}\right)+\bar{I}c_{hc}}$	$\sim -\omega$	stronger (const.)
IF	0	const.	const. (const.)
LIF	$\frac{\omega}{2\pi}\frac{\bar{I}(e^{2\pi g_L/\omega}-1)(e^{-2\pi g_L/\omega}-1)}{g_L+\bar{I}(e^{2\pi g_L/\omega}-1)}$	$\sim -\omega$	stronger (const.)

Table 3.5: Scaling of deviations in firing rate after stimulus $FL^a_{min} - FL^b$ for the different neuron models.

(3.32)-(3.33) to phase variables via (2.8), we obtain

$$d\theta_i = \left[\omega_i + z(\theta_i)I(t) + \frac{\sigma^2}{2}z(\theta_i)z'(\theta_i)\right]dt + \sigma z(\theta_i)dW_i(t) .$$
(3.34)

The $\mathcal{O}(\sigma^2)$ term is the 'Ito correction' resulting from changing variables from the stochastic differential equation (3.32) [65]. Following stimulus offset (so that $I(t) = 0$), and in the limit of small r.m.s. noise strength σ, stochastic averaging ([185],[64, Thm. 3.1]) may be applied to (3.34). As in the averaging theory of deterministic dynamical systems [82], this amounts to replacing the small deterministic term $\frac{\sigma^2}{2}z(\theta)z'(\theta)$ by its average $\tilde{z}_\omega \triangleq \frac{1}{2\pi}\int_0^{2\pi}\frac{\sigma^2}{2}z(\theta)z'(\theta)d\theta$; additionally, the coefficient of the noise term, $\sigma z(\theta)$, is replaced by the r.m.s. value $\sigma\hat{z}_\omega \triangleq \left(\frac{1}{2\pi}\int_0^{2\pi}\sigma^2 z(\theta)^2 d\theta\right)^{1/2}$, yielding

$$d\theta = \left[\omega + \tilde{z}_\omega\right]dt + \sigma\hat{z}_\omega dW(t) ,$$
(3.35)

where we have dropped the subscript i. To simplify the calculations below, we assume $\tilde{z}_\omega = 0$, as may be directly computed from the formulas in Table 2.1 for the SNIPER, Hopf, and Bautin bifurcations as well as the IF model.

The corresponding Fokker-Planck equation for the density of phase oscillators with
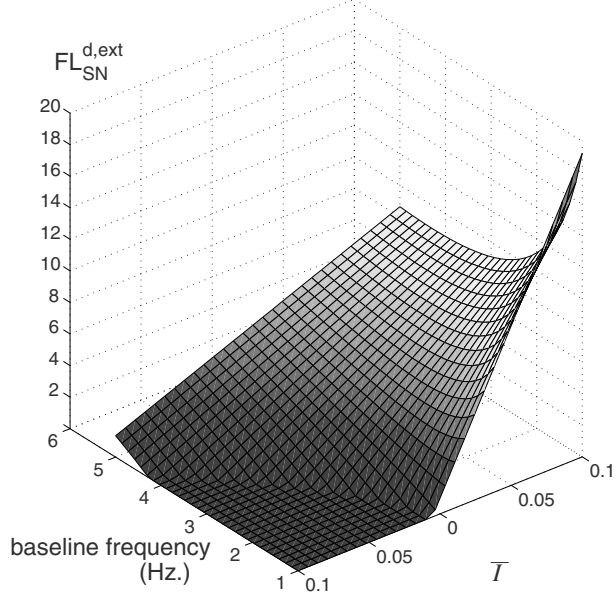
Figure 3.3: Extremal (i.e., maximally above or below baseline) firing rates $FL_{SN}^{d,ext}$ for a population of stimulated HR neurons in Hz., as a function of baseline frequency (Hz.) and applied current strength $\bar{I}$ (μA/cm^2).

baseline frequency ω is

$$\frac{\partial \rho(\theta, t, \omega)}{\partial t} = -\frac{\partial}{\partial \theta} \left[\omega \, \rho(\theta, t, \omega) \right] + \frac{\sigma^2 \hat{z_\omega}^2}{2} \frac{\partial^2}{\partial \theta^2} \left[\rho(\theta, t, \omega) \right] \qquad (3.36)$$

may then be Fourier transformed for each ω, as in [163], to yield

$$\rho(\theta, t, \omega) = \sum_{n=-\infty}^{\infty} a_n(t, \omega) \exp(in\theta) \ , \ \text{where} \ \dot{a}_n = -i\omega n a_n - \frac{\sigma^2 \hat{z_\omega}^2}{2} n^2 a_n \ . \qquad (3.37)$$

Solving the latter equations with 'initial' values $a_n(t_2, \omega)$ representing the state at stimulus end, we get:

$$FL(t, \omega) = \omega \sum_{n=-\infty}^{\infty} a_n(t_2, \omega) \exp \left[- \left(i\omega n + \frac{\sigma^2 \hat{z_\omega}^2 n^2}{2} \right) (t - t_2) \right] \ . \qquad (3.38)$$

Carrying out the average in (3.31), we obtain

$$\langle FL(t) \rangle = \int r(\omega) \omega \sum_{n=-\infty}^{\infty} a_n(t_2, \omega) \exp \left[- \left(i\omega n + \frac{\sigma^2 \hat{z}_\omega^2 n^2}{2} \right) (t - t_2) \right] d\omega \qquad (3.39)$$
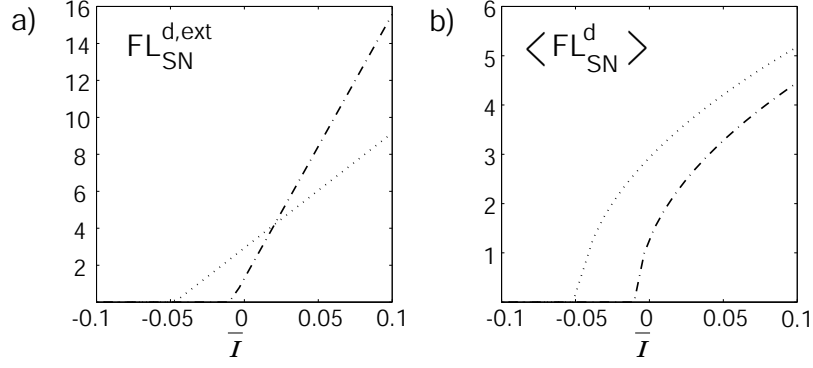
Figure 3.4: Maximum/minimum firing rate of a population of stimulated HR neurons (a), and average firing rate (b) in Hz., as a function of applied current strength for two different baseline frequencies: 1.3 Hz. (dot-dashed line), and the higher frequency 2.9 Hz (dotted). The increased gain effect at lower baseline frequencies discussed is evident for maximum/minimum, but not average, firing rates (see text).

so that

$$\langle FL(t) \rangle - \frac{\langle \omega \rangle}{2\pi} = \int_{-\infty}^{\infty} r(\omega)\, \omega \times$$

$$\sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2, \omega) \exp\left[-\left(i\omega n + \frac{\sigma^2 \hat{z}_\omega^2 n^2}{2}\right)(t - t_2)\right] d\omega , \qquad (3.40)$$

where $\langle \omega \rangle$ denotes the mean of $r(\omega)$. We will now estimate how the average spike rate $\langle FL(t) \rangle$ relaxes to its baseline value $\frac{\langle \omega \rangle}{2\pi}$.

Choosing a 'maximal' frequency ω_m beyond the essential support of the integrand, breaking the integral into pieces, and applying the triangle inequality, (3.40) yields

$$\left| FL(t) - \frac{\langle \omega \rangle}{2\pi} \right| \leq \left| \int_0^{\omega_m} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2, \omega) \exp\left[-\left(i\omega n + \frac{\sigma^2 \hat{z}_\omega^2 n^2}{2}\right)(t - t_2)\right] d\omega \right|$$

$$+ \left| \int_{\omega_m}^{\infty} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2, \omega) \exp\left[-\left(i\omega n + \frac{\sigma^2 \hat{z}_\omega^2 n^2}{2}\right)(t - t_2)\right] d\omega \right|$$

$$(3.41)$$

Noting that the exponential in the second term is bounded in modulus by 1, and that $\sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2, \omega) = FL(t_2, \omega) - \omega/2\pi$, we can bound the second term of (3.41) as

follows:

$$
\left| \int_{\omega_m}^{\infty} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-\left(i\omega n + \frac{\sigma^2 \hat{z}_\omega^2 n^2}{2} \right)(t-t_2) \right] d\omega \right| \leq
$$

$$
\int_{\omega_m}^{\infty} \left| r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-(i\omega n)(t-t_2) \right] d\omega \right| =
$$

$$
\int_{\omega_m}^{\infty} r(\omega)\omega \left| \tilde{\rho}(\theta=0, t-t_2, \omega) - \frac{1}{2\pi} \right| d\omega \leq
$$

$$
\int_{\omega_m}^{\infty} r(\omega)\omega \max_{\theta} \left| \tilde{\rho}(\theta, t_2, \omega) - \frac{1}{2\pi} \right| d\omega \triangleq E .
$$

Here, $\tilde{\rho}(\theta, t - t_2, \omega)$ is the density function that would result from evolution at frequency ω with $\sigma = 0$. Because $z_\omega(\theta) \sim 1/\omega$, the effective stimulus strength declines with ω, so that for sufficiently large ω the perturbation of $\rho(\theta, t_2, \omega)$ from equilibrium $\rho \equiv 1/2\pi$ is negligible. Because, additionally, $\omega r(\omega)$ decays for sufficiently large ω (since r is a PDF), ω_m may be chosen so that E is arbitrarily small. As for the first term in (3.41), noting that $\hat{z}_\omega$ decays with ω we can replace $\hat{z}_\omega$ in the exponential with its least value $\hat{z}_{\omega_m}$ and also replace n by 1 to obtain an upper bound on the decaying flux. This allows us to remove the exponential from the integral, giving

$$
\left| FL(t) - \frac{\langle \omega \rangle}{2\pi} \right| \leq \exp\left[-\frac{\sigma^2 \hat{z}_{\omega_m}^2 (t-t_2)}{2} \right]
$$

$$
\left| \int_0^{\omega_m} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-i\omega n(t-t_2) \right] d\omega \right| + E .
$$

$$
\tag{3.42}
$$

Next, we note that

$$
\left| \int_0^{\omega_m} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-i\omega n(t-t_2) \right] d\omega \right| =
$$

$$
\left| \int_0^{\infty} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-i\omega n(t-t_2) \right] d\omega - \right.
$$

$$
\left. \int_{\omega_m}^{\infty} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-i\omega n(t-t_2) \right] d\omega \right|
$$

$$
\leq \left| \int_0^{\infty} r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2,\omega) \exp\left[-i\omega n(t-t_2) \right] d\omega \right| + E
$$

which, together with (3.42), implies

$$\left| FL(t) - \frac{\langle \omega \rangle}{2\pi} \right| \leq \exp\left[-\frac{\sigma^2 \hat{z}_{\omega_m}^2 (t - t_2)}{2} \right] \times$$

$$\left| \int_0^\infty r(\omega)\omega \sum_{n=-\infty, n\neq 0}^{\infty} a_n(t_2, \omega) \exp\left[-i\omega n(t - t_2) \right] d\omega \right| . \qquad (3.43)$$

up to the arbitrarily small error $2E$. For each n, the integral in (3.43) is the Fourier transform of $r(\omega)\,\omega\,a_n(t_2, \omega)$ evaluated at $[n(t - t_2)]$. This integral decays for sufficiently large $t - t_2$ (by the Riemann-Lebesgue Lemma), giving an additional decay factor. Thus, (3.43) supplies our final main finding: **(VII) Response decays exponentially or faster with t due to noise and heterogeneous frequencies.**

In the special case that $r(\omega)$ is Gaussian and varies rapidly compared with $\omega\,a_n(t_2, \omega)$, for each n contributing significantly to the sum, we have

$$r(\omega)\,\omega\,a_n(t_2, \omega) \approx \frac{1}{\sqrt{2\pi\gamma}} \exp\left(-\frac{(\omega - \langle \omega \rangle)^2}{2\gamma} \right) \langle \omega \rangle\,a_n(t_2, \langle \omega \rangle) , \qquad (3.44)$$

and the integral in (3.43) may be evaluated to give the following upper bound on decay rate:

$$\left| \langle FL(t) \rangle - \frac{\langle \omega \rangle}{2\pi} \right| \leq \exp\left[-\frac{\sigma^2 \hat{z}_{\omega_m}^2 (t - t_2)}{2} - \frac{\gamma(t - t_2)^2}{2} \right] \left| \sum_{n=-\infty, n\neq 0}^{\infty} \langle \omega \rangle a_n(t_2, \langle \omega \rangle) \right|$$

$$= \exp\left[-\frac{\sigma^2 \hat{z}_{\omega_m}^2 (t - t_2)}{2} - \frac{\gamma(t - t_2)^2}{2} \right] \left| \langle FL(t_2, \langle \omega \rangle) \rangle - \frac{\langle \omega \rangle}{2\pi} \right| . \tag{3.45}$$

Here $\langle FL(t_2, \langle \omega \rangle) \rangle$ is the value of FL at time t_2 under the condition that $r(\omega) = \delta(\omega - \langle \omega \rangle)$. In Chapter 5, we study a specific application involving a (narrow) Gaussian distribution of frequencies for which (3.44) holds; and the corresponding Fig. 5.9 shown there illustrates that (3.45) provides a good decay estimate.

3.6   Applications and experimental predictions

We now provide further comments on how the mechanisms studied in this chapter could be applied and tested. As discussed in Section 3.4 and with regard to the *locus coeruleus* (LC) in the Introduction, baseline frequency-dependent variations in the sensitivity of neural populations to external stimuli could be used to adjust gain in information processing. The effect could be to engage the processing units relevant to specific tasks, and, as in [154, 167], to additionally sensitize these units to salient stimuli. See Chapter 5 and [24] for details of the LC application.

We recall that Section 3.3.2 described the different types of post-stimulus 'ringing' of firing rates $FL(t)$ that occur for the various neuron models. This 'phase-resetting'

effect has long been studied in theoretical and experimental neuroscience (e.g. [183, 163, 117]). As we show here (Eqn. (3.19), Fig. 3.2), for neuron models having a phase response curve $z(\theta)$ that takes negative values, the greatest deviations from baseline firing rates can occur significantly *after* stimulus end. Subpopulations of such neurons could be used in detecting *offsets* of sensory stimuli. Elevated firing rates $FL(t)$ that remain (or are enhanced) after the stimulus ends are an example of persistent neural activity, a general phenomenon implicated in short-term memory, interval timing, and other cognitive functions. However, physiological evidence suggests that some of the persistent activity observed *in vivo* results from desynchronized, not phase-clustered, neural groups.

Finally, stimulus-induced ringing of population firing rates (which occurs at the natural baseline frequency of the neuron population, see Eqn. (3.6)) could play a role in generating the periodic patterns of coherent activity that accompany salient events in psychological tests (e.g. 'alpha-ringing'); the possible relevance of this effect is well-known and is a topic of current debate in the EEG community [117, 14].

The results presented here are experimentally testable. As noted in the Introduction, the predictions for average firing rate $FL(t)$ are equally valid for multi-channel recordings from a (weakly coupled) population *and* for sequences of single-unit recordings from members of such a population. Thus, the $FL(t)$ predictions of this chapter can be compared with Peri-Stimulus Time Histograms (PSTHs) formed from both types of data. The scaling of response magnitudes predicted in Sect. 3.3.3 could be tested in any experiment in which baseline neural firing rates are modulated pharmacologically while stereotyped stimuli are presented. This is essentially what is done in many experiments on the effects of different neuromodulators, neurotransmitters, and other agents. For example, direct application of the neuropeptide corticotropin releasing factor (CRF) has been found to increase LC baseline activity and simultaneously decreases responses to sensory stimuli [123] in some, but not all, protocols. Many other examples of such 'modulatory' effects of neurotransmitters or exogenous inputs exist for neurons in other brain areas [8]. However, a general difficulty is that these substances may change many parameters in neurons besides the bifurcation (or frequency) parameter I_b that is the focus of this chapter, hence making it difficult to determine what mechanism leads to changes in averaged response. Furthermore, the presence of noise tends to diminish the scaling results reported here (cf. [86, 24]), and while it seems that coupling can in some circumstances amplify the scaling [24], we are still working to clarify this effect.

We close by mentioning another experimental test of the predictions presented here, suggested by John Rinzel. First, one could determine what pharmacological manipulations would cause a given *in vitro* neuron to transition from periodic firing near a SNIPER bifurcation to periodic firing near a Bautin bifurcation. Then, one could measure how trial-averaged responses to stereotyped stimuli vary as this manipulation is performed. In particular, this chapter predicts that maximal responses should occur during the stimulus in SNIPER firing, but after the stimulus switches off following a manipulation to Bautin firing.

Chapter 4

Globally coupled oscillator networks

4.1   Chapter outline

Chapter 3 was primarily concerned with the (partial, transient) synchronization of uncoupled oscillators due to external stimuli. In this chapter we consider the synchronous states of identical oscillators, with differing phase relationships, that arise due to coupling.

   In particular, we study a class of permutation-symmetric, globally-coupled phase oscillator networks on N-dimensional tori. Following the introductory material of Section 4.2, in Section 4.3 we focus on the effects of rotation and reflection symmetries and of the spectral content of coupling functions on the existence (Sections 4.3.1-4.3.4), stability, and degeneracy (Sections 4.3.5-4.3.7) of phase-locked solutions in which subgroups of oscillators share common phases. We also estimate domains of attraction for the completely synchronized state. Richer coupling functions, which break some of the symmetry exploited in Section 4.3, are introduced in Section 4.4. In Section 4.5, the results are applied to coupling functions derived from the Hodgkin-Huxley conductance-based neuron model (i.e., Eqns. (2.1-2.2)). In particular, implications for stochastically forced networks are illustrated numerically via cross-correlograms of spike times, and effects of phase-difference- and individual-phase-dependent coupling on the frequency of the fully synchronized state are briefly discussed. Many of the results in this Chapter were published in [22].

4.2   Introduction and background

We consider networks of N rotator oscillators with constant forcing and pairwise phase-difference and absolute-phase 'product' coupling, described by:

$$\dot{\theta}_i = \omega_i + \frac{1}{N} \sum_{j=1}^{N} \alpha_{ij} f_{ij}(\theta_j - \theta_i) + h_i(\theta_i) \frac{1}{N-1} \sum_{j \neq i}^{N} \beta_{ij} g_j(\theta_j), \qquad (4.1)$$

where $(\theta_1, \ldots, \theta_N)^{\mathrm{T}} \in \mathrm{T}^N$, α_{ij}, β_{ij} and f_{ij}, h_i, g_j are, respectively, coupling parameters and 2π-periodic functions, and ω_i are the natural frequencies of the uncoupled rotators. This chapter focuses on networks with identical frequencies and global (mean field) coupling, so that equation (4.1) becomes

$$\dot{\theta}_i = \omega + \frac{\alpha}{N} \sum_{j=1}^{N} f(\theta_j - \theta_i) + h(\theta_i) \frac{\beta}{N-1} \sum_{j \neq i}^{N} g(\theta_j) , \qquad (4.2)$$

although we include some results with additive random noise. The denominators (N, $N-1$) are introduced to normalize coupling effects.

Rotator (phase-only) models of coupled oscillators have been widely studied, especially in the contexts of neuroscience and coupled Josephson junctions. As seen in Chapter 2, the phase equations offer, respectively, significant simplification of more realistic neuron models of Hodgkin-Huxley or Fitzhugh-Nagumo type (see also [126, 102, 94]), and of the Josephson circuit equations: e.g. [173, 174, 178]. As also noted in Chapter 2, in the case that the N uncoupled oscillators have strongly attracting limit cycles in their full phase space, the persistence of normally hyperbolic invariant manifolds [59] under small perturbations (weak coupling) may be used to reduce the system to the N-torus by a suitable coordinate transformation. The distinct 'strong attraction' (SA) [53, 94] and 'phase response' [50, 111] (PR) techniques for approximating the reduced system (discussed in Section 2.2.3 of Chapter 2) will be applied, to differing results, in Section 4.5 of this chapter.

In motivating Eqn. (4.2), we assume an additional separation of scales between the "electrotonic" and "synaptic" types of neural coupling (introduced in Section 4.5). We take electrotonic coupling to be weaker than synaptic, so that it can be averaged to give the phase-difference functions αf_{ij} and assume that this does not affect the $\beta h_i g_j$ terms at leading order. Sections 4.3 and 4.4 consider the dynamics of equation (4.1) for various values of α and β, without a priori restricting to the $|\alpha| \ll |\beta| \ll \mathcal{O}(1)$ required in this derivation of the phase equations.

When $\beta = 0$ but frequencies differ between oscillators, equation (4.2) is referred to as the Kuramoto model ([110]), on which there is an extensive literature; see the recent review of [162] and references therein (e.g. [42]). Much of this work has been done in the continuum limit $N \to \infty$, and [162] adopts this viewpoint; specifically, stability analyses of some stationary (continuous) states are discussed. Finite-dimensional results, including a Liapunov function and dimension reduction, are found in the context of Josephson junction models in [173]. Many earlier studies take only the leading term in an odd Fourier expansion of f, so that $f(\cdot) = \sin(\cdot)$; as we shall see this is a very degenerate case for the mean field coupled system (4.2) (e.g. [128, 75]). Moreover, as shown in [97], relaxation oscillators of Hodgkin-Huxley or Fitzhugh-Nagumo type lead to much richer phase difference functions than $\sin(\cdot)$. Others have recognized the importance of higher Fourier harmonics: see [44, 75, 128, 174]. Additional work on finite dimensional oscillator networks includes [106, 108, 107], which consider directed coupling, [17], which considers integrate-and-fire models derived from coupled spiking neurons, and [131], which will be discussed in

Section 4.3. Recently, [32] has studied the effects of spike shape on electrotonically coupled integrate-and-fire networks. The authors find that the existence and stability of splay states depends on the spike shape in a manner that would be interesting to compare with the present results.

The present chapter draws on [6], which addresses a class of $S_N \times T^1$-equivariant oscillator networks (of which (4.2) is an example when $\beta = 0$). We now summarize the properties of symmetric dynamical systems necessary to present and apply these results; for more background, see [76, 77].

Consider the ODE

$$\frac{dx}{dt} = G(x) \ , \ x \in \text{manifold } M \ , \tag{4.3}$$

and let Γ be a group acting on M. The ODE is said to be Γ-equivariant if f commutes with the group action, i.e.

$$G(\gamma x) = \widehat{\gamma} G(x) \ \ \forall \gamma \in \Gamma, x \in M \ , \tag{4.4}$$

where the derivative map $\widehat{\gamma}$ ([4]) acts on the tangent space TM; for linear actions of γ, $\widehat{\gamma} = \gamma$. The symmetry of a solution $x_0 \in M$ is characterized by the isotropy subgroup $\Sigma_{x_0} = \{\gamma \in \Gamma : \gamma x_0 = x_0\}$, that is, the set of all group elements which leave the solution x_0 unchanged. Associated with an isotropy subgroup is a fixed point subspace $\text{Fix}[\Sigma_{x_0}] = \{x \in M : \sigma x = x \ \ \forall \sigma \in \Sigma_{x_0}\}$: the set of points fixed by all elements of Σ_{x_0}. Two immediate consequences of Γ-equivariance are that (1) for any solution $x(t)$ to equation (4.3), $\gamma x(t)$ is also a solution, and (2) fixed-point subspaces are invariant under the flow generated by G. We will refer to this latter property as dynamical invariance. As in [6], we study special classes of symmetric systems defined by the following groups: the circle group $T^1 = \{\delta : \ \delta \in [0, 2\pi)\}$ (with action on T^N, $\theta_i \mapsto \theta_i + \delta, \ \forall \ i$), the cyclic subgroups $Z_m \in T^1$ (with action $\theta_i \mapsto \theta_i + 2\pi/m$), the subgroups of permutations on j-many coordinates, S_j, and the reflection group Z_2^r with action $\theta_i \mapsto -\theta_i, \ \forall \ i$.

The remainder of the chapter proceeds as follows. In Section 4.3 we study (4.2) with $\beta = 0$ ($S_N \times T^1$ equivariant), emphasizing the influence of general coupling functions and obtaining additional results for odd functions f. In Section 4.4 the T^1 symmetry is broken through re-introduction of $h(\theta_i)g(\theta_j)$ terms. Finally, Section 4.5 applies the results of previous sections to two different phase reductions of the Hodgkin-Huxley equations with electrotonic and synaptic neural coupling. Thus, Sections 4.3 and 4.4 are largely abstract and general, while Section 4.5 concerns a specific neural application.

4.3 Existence and stability of phase locked solutions for systems with a rotation symmetry

In this section we study dynamical systems of the general form

$$\dot{\theta} = \omega + G(\theta) \ ; \ \theta \in \text{T}^N \tag{4.5}$$

equivariant under rotation and permutation (actions of $S_N \times T^1$), or equivariant under rotation, permutation, *and* reflection ($S_N \times T^1 \dot{+} Z_2^r$). In both cases we immediately transform to coordinates $\phi_j = \theta_j - \omega t$ rotating with the natural frequency, so that (4.5) becomes

$$\dot{\phi} = G(\phi) \; ; \; \phi \in T^N \; , \tag{4.6}$$

and seek phase locked solutions, in which all phase differences $\phi_j - \phi_i \equiv \phi_{ji}$ are constant. In particular, we look for 'diagonal flow' periodic solutions $\bar{\phi}$ such that

$$\dot{\bar{\phi}}_i \equiv c, \quad i = 1, \ldots, N, \tag{4.7}$$

where c is a constant, nonzero in general. These solutions are also periodic for (4.5), and, employing a second rotating frame $\theta_i - (\omega + c)t$, they become fixed points.

Sections 4.3.1 and 4.3.3 obtain solutions for general systems of the form (4.6), using tools from equivariant dynamics. In Sections 4.3.2 and 4.3.4, we consider the special case of (4.6) in which coupling is of a pairwise phase-difference form, so that

$$\dot{\phi}_i = \frac{\alpha}{N} \sum_{j=1}^{N} f\left(\phi_j - \phi_i\right), \quad i = 1, \ldots, N, \tag{4.8}$$

and thereby obtain additional results as well as simplified proofs. As discussed above, this type of coupling corresponds to generic weakly-coupled oscillator systems. The function $f(\cdot)$ is assumed to be continuously differentiable and 2π-periodic. To state some of the results below, it is useful to express f in a Fourier series with coefficients b_l^o and b_l^e:

$$f(\phi_{ji}) = \sum_{l=0}^{\infty} \left(b_l^o \sin(l\phi_{ji}) + b_l^e \cos(l\phi_{ji})\right). \tag{4.9}$$

While several of the following results appear in the literature as noted below, no unified presentation seems to exist, so we provide a summary here, including extensions and new examples of our own.

4.3.1 Existence of phase locked solutions for general $S_N \times T^1$ equivariant systems

We now discuss isotropy subgroups: subgroups of the 'original' equivariant group $S_N \times T^1$ which fix a (group orbit of) linear subspace(s) of T^N (by definition, their fixed point subspaces). Recall that the importance of this lies in the fact that fixed point subspaces of isotropy subgroups are dynamically invariant.

Ashwin and Swift [6] found all of the isotropy subgroups of $S_N \times T^1$ acting on T^N. To understand what these are, it is easiest to study configurations of phases which comprise their fixed point subspaces. The first step is to partition the N phases ϕ_i into m blocks, each containing k phases. Then, further partition each of these m blocks into l_B subblocks each containing k_j phases ($k = (k_1 + \cdots + k_{l_B})$). See Fig. 4.1 (a). Let $S_{k_j} \subset S_N$ be the subgroup of permutations of phases within each subblock
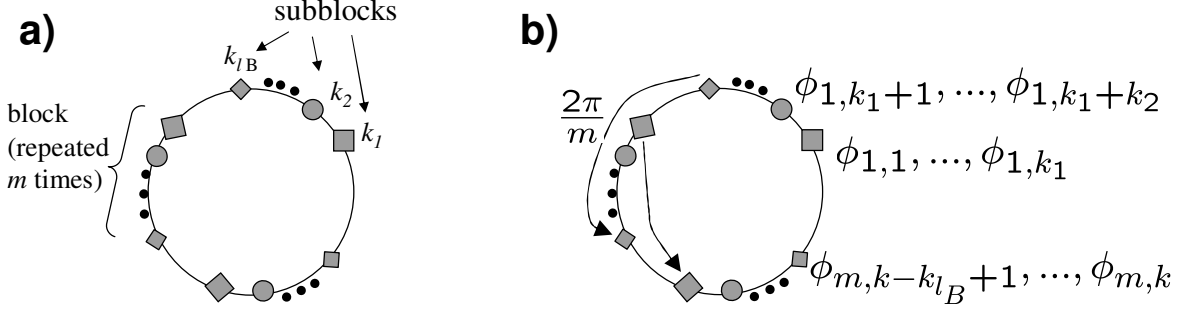
Figure 4.1: A general element of $Fix\left[(S_{k_1} \times \cdots \times S_{k_{l_B}})^m \dot{+} Z_m(\rho_m)\right]$, shown twice to emphasize different features. (a) Partition of oscillators among m blocks and k_{l_B} subblocks. (b) Labelling of phases. Action of $Z_m(\rho_m)$ is rotation by $\frac{2\pi}{m}$ (shown) followed by relabelling of phases (see text).

of k_j phases. Then the subgroup $(S_{k_1} \times \cdots \times S_{k_{l_B}})^m \subset S_N$ leaves our partition of oscillators invariant, no matter how the blocks and subblocks of phases are arranged around the circle.

We now address spacing among the subblocks. Consider the cyclic subgroups $Z_m \subset T^1$ (with action generated by $\phi_i \mapsto \phi_i + 2\pi/m$). If the m blocks are regularly spaced as in Fig. 4.1, the resulting configuration will be invariant under the additional subgroup $Z_m(\rho_m) \subset S_N \times T^1$, which, following [6], is defined as follows. Let $l = 1, ..., m$ index the blocks in order of increasing phase, and fix some ordering of the k phases in the lth of the m blocks as $\phi_{l,1}, ..., \phi_{l,k}$, where $\phi_{l,1}, ..., \phi_{l,k_1}$ belong to the 1st subblock of the lth block, $\phi_{l,k_1+1}, ..., \phi_{l,k_1+k_2}$ belong to the 2nd, etc. See Fig. 4.1 (b). Then the generator of the m-element subgroup $Z_m(\rho_m) \subset S_N \times T^1$ acts via the shift $\phi_i \mapsto \phi_i + 2\pi/m$, followed by the permutations $(\phi_{m,1}, \phi_{(m-1),1}, ..., \phi_{1,1})$, $(\phi_{m,2}, \phi_{(m-1),2}, ..., \phi_{1,2})$, ..., $(\phi_{m,k}, \phi_{(m-1),k}, ..., \phi_{1,k})$, which 'relabel' the rotated oscillators between sequentially ordered blocks.

Taking the (semi) direct products of the two subgroups discussed above gives (exactly) the whole family of isotropy subgroups:

Theorem 4.3.1. *(Ashwin and Swift [6]) Every isotropy subgroup of a general $S_N \times T^1$-equivariant vector field is of the form:*

$$\Sigma_{\mathbf{k},m} \equiv (S_{k_1} \times \cdots \times S_{k_{l_B}})^m \dot{+} Z_m(\rho_m) ,$$

where $N = m(k_1 + \cdots + k_{l_B})$.

The permutations all commute, hence the direct products, while the Z_m action does not commute with the permutations, hence the semi-direct product denoted by $\dot{+}$. See Fig. 4.2 for examples.

Overall, the fixed-point subspace Fix[$\Sigma_{\mathbf{k},m}$] is an l_B-torus: there are $l_B - 1$ degrees of freedom setting the spacings between the blocks, plus an additional degree of freedom determining a 'reference' ϕ_1; this represents the T^1 group orbit.
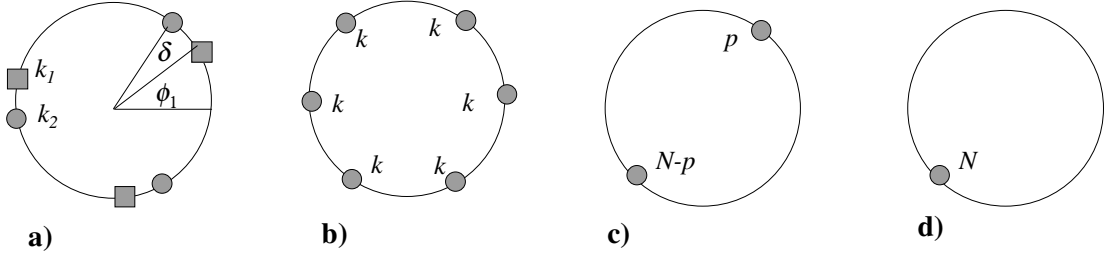
Figure 4.2: Illustrations of equilibria fixed under the actions of various isotropy subgroups. (a) An element of $\text{Fix}[(S_{k_1} \times S_{k_2})^3 \dot{+} Z_3]$. Each square represents k_1 oscillators mutually in phase, and successive squares denote groups differing in phase by $2\pi/3$ (similar for circles). Elements of this fixed point subspace are parameterized by two angles ϕ_i and δ, so it is a 2-torus. (b) $(S_k)^m \dot{+} Z_m$ (rotating block modes) with $m = 6$, (c) $S_p \times S_{N-p}$ (two-block modes), and (d) S_N (in-phase mode).

Ashwin and Swift then show that there exist periodic orbits with diagonal flow having *maximal* symmetry given by certain of these isotropy subgroups:

Theorem 4.3.2. *(Ashwin and Swift [6]) $S_N \times T^1$ equivariant flows on the N-torus generically have periodic orbits of diagonal flow with* maximal *isotropy $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$, for each m, $k_1 \neq k_2$ such that $m(k_1 + k_2) = N$.*

This theorem is proved by noting that, without loss of generality, the phases of the oscillators can be ordered as $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_N \leq \phi_1 + 2\pi$. The oscillators retain their ordering under the dynamics, i.e., they can never 'pass' each other, since this would involve crossing an invariant fixed-point subspace. Projecting the phases onto the manifold $\phi_1 = 0$ (by subtracting the instantaneous value of ϕ_1 from each phase) gives a simplex called the 'canonical invariant region' (CIR). The intersection of $\text{Fix}[(S_k)^m \dot{+} Z_m]$ with the CIR is a zero-dimensional invariant subspace, i.e., an equilibrium. In the unprojected system, this corresponds to a periodic orbit (or a circle of equilibria if $\phi_1 = 0$ for all time) with isotropy $(S_k)^m \dot{+} Z_m$. Furthermore, the intersection of $\text{Fix}[(S_{k_1} \times S_{k_2})^m \dot{+} Z_m]$ with the CIR is a one-dimensional line segment. The end points of this segment have isotropy $(S_{k_1+k_2})^m \dot{+} Z_m$, and are equilibria with stability in the direction of the line segment determined by the same eigenvalue. Provided this eigenvalue does not vanish (this is the nondegeneracy condition satisfied for generic functions f), this can only happen if there is at least one equilibrium in the interior of the line segment. In the original system, this corresponds to a periodic orbit (or one-torus of equilibria if $\dot{\phi}_1 = 0$ for all time) with isotropy $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$. If $k_1 = k_2$, the midpoint of the line segment is an equilibrium with isotropy $(S_{k_1})^{2m} \dot{+} Z_{2m}$; this can serve as the necessary equilibrium in the interior of the line segment. With an additional nondegeneracy condition, Ashwin and Swift's result can be extended to show that periodic orbits of diagonal flow exist with maximal isotropy $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$ with $k_1 = k_2$ as well.

4.3.2 Additional results and simplified proofs for pairwise phase-difference $S_N \times T^1$ equivariant systems

As discussed in the previous section, Ashwin and Swift [6] developed their existence proof for systems coupled with general T^1-equivariant functions (4.6). The special additive, pairwise-coupled form of the coupling in (4.8), which we study in this section, allows a much simpler argument to prove (a restricted version of) Theorem 4.3.2.

For a $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$ solution with each pair of clusters separated by phase δ (see Figure 4.2(a)) to have diagonal flow requires $\dot{\phi}_i \equiv c(\delta) \; \forall i$, for some fixed δ. This condition reduces to

$$c_1(\delta) = c_2(\delta) \; , \; \text{where} \tag{4.10}$$

$$c_1(\delta) = k_1 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m}\right) + k_2 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m} + \delta\right) \tag{4.11}$$

$$c_2(\delta) = k_2 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m}\right) + k_1 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m} - \delta\right) \tag{4.12}$$

are the (constant) phase velocities for oscillators in k_1 or k_2-clusters, respectively (cf. [103] for $m = 1$). A quick sketch shows that at least one $\delta \in (0, 2\pi/m)$ satisfying (4.10) must exist if $c_1'(0)/k_2 = -c_2'(0)/k_1$ is nonzero, since $c_{1,2}'(0) = c_{1,2}'(2\pi/m)$. Thus, the nondegeneracy condition becomes $\sum_{j=0}^{m-1} f'(\frac{2\pi j}{m}) \neq 0$; for $S_{N-p} \times S_p$ solutions, $m = 1$, implying $f'(0) \neq 0$. Further, for $(S_k)^m \dot{+} Z_m$ rotating blocks the equality $\dot{\phi}_i \equiv \gamma$ is automatic. Finally, we note that if $k_1 = k_2$, $\delta = \pi/m$ always satisfies (4.10), so that the corresponding $(S_{k_1} \times S_{k_1})^m \dot{+} Z_m$ solutions may also have symmetry $(S_{k_1})^{2m} \dot{+} Z_{2m}$.

These arguments extend in a natural way to show the existence of weak solutions to the partial differential equations derived from (4.8) as $N \to \infty$ (see [43] for the derivation of this PDE). These are symmetrically-spaced combinations of delta distributions rotating at the frequency $c(\delta)$ found above, with the k_j-cluster distributions weighted by $k_j(N)/m(k_1(N) + k_2(N))$, $j = 1, 2$. Here, $k_j(N)$ is the number of oscillators in a cluster when the total number of oscillators is N, and the $N \to \infty$ limit is taken over a subsequence of configurations ([45]) with constant $k_j(N)/m(k_1(N) + k_2(N))$ such that m (fixed) divides $k_1(N) + k_2(N)$. Under the same nondegeneracy conditions as above, their existence may be shown for *any* values of $k_j/m(k_1 + k_2)$ and *any* m. Furthermore, if f lacks m-th Fourier harmonics and their multiples, families of Z_m-symmetric solutions analogous to the fixed tori of Section 4.3.7 also exist.

Despite the variety of these equilibria, we can prove a result on the *non*-existence of fixed points in a region surrounding the in-phase solution. Define the open N-cylinder $\mathcal{C}_{R_1} \stackrel{\triangle}{=} \{\theta | d(\theta, \theta_d(\psi)) \leq R_1 \text{ for some } \psi \in [0, 2\pi]\}$. Here, $d(\cdot, \cdot)$ is the Euclidian metric on $\mathcal{R}^N$ (and hence on T^N) and $\theta_d(\psi)$ is the N-vector with all coordinates equal to ψ (so that the axis of $\mathcal{C}_{R_1}$ is the diagonal $\mathcal{D} \stackrel{\triangle}{=} \{\theta | \theta_i = \theta_j \; \forall \; i, j\}$; see Figure 4.3).

Proposition 4.3.1. *Let $R_1 > 0$ be such that* either *f restricted to $(0, 2R_1)$ or f restricted to $(-2R_1, 0)$ is of one sign (i.e., f is strictly negative or positive in the*
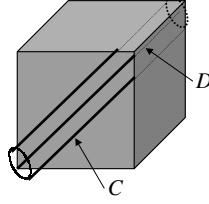
Figure 4.3: The diagonal $\mathcal{D}$ and cylinder $\mathcal{C}$ of Proposition 4.3.1. The cube represents the N-torus.

region). Then there are no fixed points of (4.8) in $\mathcal{C}_{R_1} \setminus \mathcal{D}$.

Proof. First we note that for $\phi \in \mathcal{C}_{R_1}$, there exists a ψ s.t. $\phi \in B_{R_1}(\psi) \stackrel{\triangle}{=} \{\phi | d(\phi, \theta_d(\psi)) < R_1)\}$. Thus in particular $|\phi_j - \psi| < R_1 \; \forall j$; summing two of these inequalities and applying the triangle inequality gives $|\phi_{ji}| < 2R_1 \; \forall \; i, j$. Next, consider an arbitrary N-vector $\phi \in \mathcal{C}_{R_1} \setminus \mathcal{D}$, where without loss of generality $\phi_1 \leq \phi_2 \leq \cdots \leq \phi_N$. If the chosen interval in the hypothesis is $(0, 2R_1)$, note that $2R_1 > \phi_j - \phi_1 \geq 0 \;\; \forall \; j$ and $\phi_j - \phi_1 > 0$ for at least one j (since $\phi \notin \mathcal{D}$). Thus, $\dot{\phi}_1 = \frac{\alpha}{N} \sum_{j=1}^{N} f(\phi_j - \phi_1) \neq 0$, because each term in the sum is either zero or of the same sign (the continuity of f implies that $f(0)$ is either 0 or of the same sign as $f(\phi)$ for $\phi \in (0, 2R_1)$) and at least one term is nonzero. If the chosen interval is $(-2R_1, 0)$, we use the facts that $0 \geq \phi_j - \phi_N > -2R_1 \;\; \forall \; j$ and $\phi_j - \phi_N < 0$ for at least one j. Similarly, then, $\dot{\phi}_N = \frac{\alpha}{N} \sum_{i=1}^{N} f(\phi_j - \phi_N) \neq 0$. Hence ϕ is not a fixed point. $\square$

This result does not exclude orbits with nonzero diagonal flow, which may exist within $\mathcal{C}_{R_1} \setminus \mathcal{D}$, but only if f is *not* odd (see below).

4.3.3 Existence of phase locked solutions for general $S_N \times T^1 \dot{+} Z_2^r$ equivariant systems

We now consider systems with an additional reflection symmetry, e.g. in the case that f is an odd function, so that the complete symmetry group is $S_N \times T^1 \dot{+} Z_2^r$; recall that the nontrivial action of the two-element group Z_2^r on T^N is the reflection $\theta_i \mapsto -\theta_i \; \forall \; i$. For the general system (4.6), Z_2^r equivariance implies $G(-\theta) = -G(\theta)$.

We now define the isotropy subgroups of $S_N \times T^1 \dot{+} Z_2^r$. The set of these subgroups obviously contains those identified in Thm. 4.3.1. We ask if more isotropy subgroups may be found in the presence of the additional Z_2^r symmetry. Our strategy is to construct configurations invariant under various subgroups of $S_N \times T^1 \dot{+} Z_2^r$.

We begin with the same initial partition as in the previous section: break the N phases into m blocks, each containing k phases. We then partition each of these m blocks into $2l_B + 1$ subblocks, where subblocks $j, j + 1$ each contain k_j phases, $j = 1, ..., l_B$, and the final subblock ($j = 2l_B + 1$) has k_{l_B+1} phases ($k = 2k_1 + \cdots + 2k_{l_B} + k_{l_B+1}$). See Fig. 4.4(a). The subgroup $(S_{k_1} \times S_{k_1} \times \cdots \times S_{k_{l_B}} \times S_{k_{l_B}} \times S_{k_{l_B}+1})^m \subset S_N$
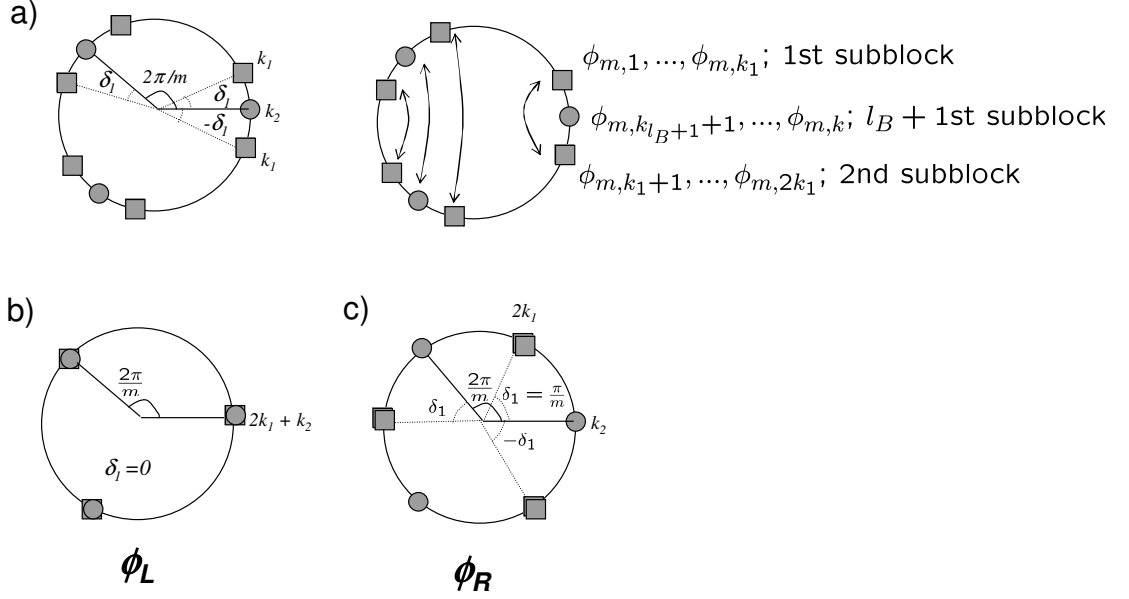
Figure 4.4: (a) A general element ϕ of $Fix\left[(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)\right]$ for $m = 3$, shown twice to highlight different features. Left, partition and spacing of phases. Right, phase labelling and action of the Z_2^r flip $\phi_j \to -\phi_j$, which is undone by the permutation $\rho_2^r \in S_N$. This fixed point subspace is 2 dimensional in the original coordinates, the degrees of freedom being δ_1 and a rigid rotation. (b) The special "left" endpoint element of ϕ_L of $Fix\left[(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)\right]$ obtained by setting $\delta_1 = 0$. (c) The other endpoint ϕ_R, with $\delta_1 = \pi/m$.

leaves our partition of oscillators invariant. Additionally, if the m blocks are evenly distributed as in Fig. 4.4(a), the configuration will be invariant under

$$\Sigma_{\mathbf{k},m} \equiv (S_{k_1} \times S_{k_1} \times \cdots \times S_{k_{l_B}} \times S_{k_{l_B}} \times S_{k_{l_B}+1})^m \dot{+} Z_m(\rho_m) \ ,$$

exactly as in Section 4.3.1.

If we make an additional restriction on the configuration so that pairs of subblocks are arranged on either side of a 'central point' (to be explained), the configuration will be fixed under an additional subgroup of $S_N \times T^1 \dot{+} Z_2^r$. Without loss of generality (i.e., up to rigid rotation due to T^1 symmetry), we define our configuration as follows. Let $\frac{2\pi l}{m}$ be the center for the lth block, meaning that the (l_B+1)th subblock of the lth block will have phase $\frac{2\pi l}{m}$ and that the $(j+1)$st subblocks will have phase $\frac{2\pi l}{m} - \delta_j$ and $\frac{2\pi l}{m} + \delta_j$, respectively, $j = 1, 3, 5, ..., 2l_B - 1$. See Fig. 4.4(a) for the example $l_B = 1$, $m = 3$. In what follows, we refer to a general configuration of this type simply as ϕ.

Next, motivated by [5], define the two-element subgroup $Z_2^r(\rho_2^r) \subset S_N \times T^1 \dot{+} Z_2^r$ as follows. Fix some ordering of the k phases in the lth of the m blocks as $\phi_{l,1}, ..., \phi_{l,k}$, where, as in Section 4.3.1, $\phi_{l,1}, ..., \phi_{l,k_1}$ belong to the 1st subblock of the lth block, etc. Then (the nontrivial element of) $Z_2^r(\rho_2^r)$ acts via the flip $\phi_i \mapsto -\phi_i \ \forall i$, followed by the

permutations $(\phi_{l,j}\, ,\, \phi_{(m-l),(j+1)})$, for all combinations of subblock indices $j = 1, ..., l_B$ and block indices $l = 1, ..., \lfloor m/2, \rfloor$; $(\phi_{l,j}\, ,\, \phi_{(m-l),j})$ for subblock index $j = l_B + 1$ and block indices $l = 1, ..., \lfloor m/2, \rfloor$; and $(\phi_{l,j}\, ,\, \phi_{l,(j+1)})$ for subblock indices $j = 1, ..., l_B$ and block index $l = m$.

Now, note that general configurations of the form of "ϕ" introduced in the previous configurations are fixed by

$$T_{\mathbf{k},m} \equiv (S_{k_1} \times S_{k_1} \times \cdots \times S_{k_{l_B}} \times S_{k_{l_B}} \times S_{k_{l_B}+1})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r) \ ,$$

so that the $T_{\mathbf{k},m}$ are isotropy subgroups of $S_N \times T^1 \dot{+} Z_2^r$. Therefore we have identified new isotropy subgroups:

Proposition 4.3.1. *Isotropy subgroups of $S_N \times T^1 \dot{+} Z_2^r$ include those of the form:*

$$\Sigma_{\mathbf{k},m} \equiv (S_{k_1} \times \cdots \times S_{k_{l_B}})^m \dot{+} Z_m(\rho_m) \ ,$$

where $N = m(k_1 + \cdots + k_{l_B})$
and of the form

$$T_{\mathbf{k},m} \equiv (S_{k_1} \times S_{k_1} \times \cdots \times S_{k_{l_B}} \times S_{k_{l_B}} \times S_{k_{l_B}+1})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)$$

where $N = m(2k_1 + \cdots + 2k_{l_B} + k_{l_B+1})$.

We conjecture that Proposition 4.3.1 actually lists all of the isotropy subgroups (up to conjugacy) of $S_N \times T^1 \dot{+} Z_2^r$.

In analogy with Theorem 4.3.2, we now identify particular isotropy subgroups which give the maximal symmetry of diagonal flow-periodic solutions. In this case an inequality will distinguish the open sets of oscillator systems which are guaranteed to have periodic solutions with the desired maximal isotropy; this is a difference from Theorem 4.3.2, in which the solutions of interest existed for generic systems.

Proposition 4.3.2. *If the eigenvalues λ_L and λ_R defined in the proof below are nonzero and of the same sign, then the $S_N \times T^1 \dot{+} Z_2^r$ equivariant flow (4.8) on the N-torus has a periodic orbit of diagonal flow with* maximal *isotropy $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)$, where m, k_1, and k_2 are such that $m(2k_1 + k_2) = N$.*

An element of the fixed point subspace $Fix\left[(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)\right]$ is shown in Fig. 4.4.

Proof. The proof follows the pattern of that given by Ashwin and Swift [6] for Theorem 4.3.2 above. The two dimensional dynamically invariant space $Fix\left[(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)\right]$ has a one dimensional intersection with the space defined by $\phi_1 = 0$. Therefore, transforming to coordinates rotating with ϕ_1 and w.l.o.g. setting $\phi_1 = 0$ we obtain in these new coordinates a one dimensional invariant space of configurations of the form shown in Fig. 4.4 and parameterized by δ_1.

This one dimensional space is itself split into invariant line segments defined by ordering of the subblocks; the canonical invariant region in which phases are arranged

in increasing order (discussed above in the context of Theorem 4.3.2) corresponds to a segment with endpoint values of $\delta_L = 0$, $\delta_R = \frac{\pi}{m}$ at which certain subblocks coincide (the subscripts here stand for left and right). See Fig. 4.4. The configurations ϕ_L and ϕ_R at these left and right endpoints have isotropy $(S_{2k_1+k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)$ and $(S_{2k_1} \times S_{k_2})^m \dot{+} Z_m \dot{+} Z_2^r$, respectively. Note that, since they have different isotropy, the the stability of these two phase locked states in the direction of the invariant line segment (parameterized by δ_1) connecting them is *not* identical, unlike the case of Theorem 4.3.2.

In particular let λ_L be the eigenvalue corresponding to the eigenvector of the Jacobian of (4.6) at ϕ_L pointing toward ϕ_R and λ_R be the eigenvalue for same eigenvector of the Jacobian at ϕ_R. If λ_L and λ_R are nonzero and of the same sign then there must be least one fixed point in the interior of the line segment. In the original coordinate system, this point corresponds to a periodic orbit (or one-torus of equilibria if $\dot{\phi}_1 = 0$ for all time) with maximal isotropy $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)$. $\qquad\square$

4.3.4 Additional results and simplified proofs for pairwise phase-difference $S_N \times T^1 \dot{+} Z_2^r$ equivariant systems

For the case of pairwise phase-difference coupling (4.8), Z_2^r equivariance implies $f(\theta_j - \theta_i) = -f(\theta_i - \theta_j)$; in other words, coupling functions $f(\cdot)$ must be odd.

Existence of phase-locked solutions

We may use the phase-difference form of (4.8) to explicitly calculate the eigenvalues λ_L and λ_R of Proposition 4.3.2, yielding the companion result:

Proposition 4.3.3. *If*

$$\lambda_L \equiv -(2k_1 + k_2) \sum_{j=0}^{m-1} f'\left(\frac{2\pi j}{m}\right) = -(2k_1 + k_2)m \sum_{l \in \mathcal{M}(m)} l b_l^o \qquad (4.13)$$

and

$$\lambda_R \equiv -2k_1 \sum_{j=0}^{m-1} f'\left(\frac{2\pi j}{m}\right) - k_2 \sum_{j=0}^{m-1} f'\left(\frac{2\pi j}{m} + \frac{\pi}{m}\right) \qquad (4.14)$$

$$= -2k_1 m \sum_{l \in \mathcal{M}(m)} l b_l^o - k_2 m \sum_{l \in \mathcal{M}(m)} l b_l^o \cos(\frac{l\pi}{m}) \qquad (4.15)$$

are nonzero and of the same sign, then the system (4.8) has a periodic orbit of diagonal flow with maximal *isotropy* $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m(\rho_m) \dot{+} Z_2^r(\rho_2^r)$, *where* m, k_1, *and* k_2 *such that* $m(2k_1 + k_2) = N$. *A sufficient condition for this to occur is*

$$k_1 > \frac{1}{2} k_2 \left| \frac{\sum_{l=0}^{m-1} f'\left(\frac{2\pi l}{m} + \frac{\pi}{m}\right)}{\sum_{l=0}^{m-1} f'\left(\frac{2\pi l}{m}\right)} \right| \quad . \qquad (4.16)$$
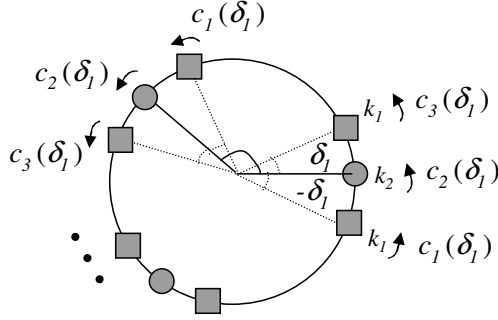
Figure 4.5: Phase velocities (i.e. values of $\dot{\phi}_j$) for the different subblocks; labeling as used in Remark 4.3.1.
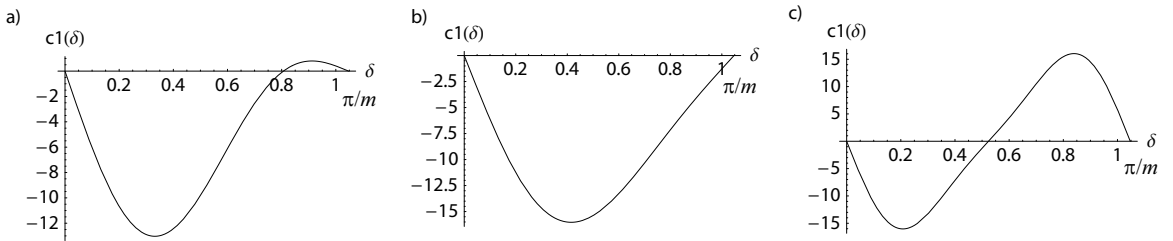
Figure 4.6: The phase velocity $c_1(\delta_1)$ of Remark 4.3.1 Examples 1 and 2. (a) for $m = 3$ and $f(\cdot) = \sin(3\cdot) + \sin(2\cdot)$, a $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m \dot{+} Z_2^r$ solution exists for $k_1 = 2$, $k_2 = 3$. (b) However, no such solution exists for $k_1 = 1$, $k_2 = 5$. (c) For $m = 3$ again but $f(\cdot) = \sin(6\cdot) + \sin(2\cdot)$, all combinations of k_1 and k_2 give $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dot{+} Z_m \dot{+} Z_2^r$ phase locked solutions; plotted, $k_1 = 1$, $k_2 = 5$.

Proof. This result follows from direct calculation. The relevant eigenvector is

$$v_\delta = (\underbrace{-1, ..., -1}_{k_1}, \underbrace{1, ..., 1}_{k_1}, \underbrace{0, ..., 0}_{k_2}, \underbrace{...}_{}, \underbrace{-1, ..., -1}_{k_1}, \underbrace{1, ..., 1}_{k_1}, \underbrace{0, ..., 0}_{k_2}) \qquad (4.17)$$

$$\underbrace{\hspace{9cm}}_{\text{(repeated m times)}}$$

which reflects variations in δ_1: the central blocks of k_2 oscillators remain unchanged, while the distance δ_1 separating them from the flanking blocks of k_1 oscillators is contracted or expanded. Multiplying by the Jacobian matrix $J(\phi_L)$ gives $J(\phi_L)v_\delta = \lambda_L v_\delta$, with λ_L as given in the Proposition; likewise, $J(\phi_R)v_\delta = \lambda_R v_\delta$ (see Fig. 4.4 for illustrations of ϕ_L and ϕ_R). $\qquad \square$

Remark 4.3.1. *In direct analogy to Section 4.3.2 above, we can write down the phase velocities $c_1(\delta_1)$ and $c_2(\delta_1)$ of the first (i.e., with least-positive phase) and second k_1*

61

sublocks and the phase velocity $c_3(\delta_1)$ of the k_2 subblocks. See Fig. 4.5. Our objective, as above, is to check under what conditions there exists a value of $\bar{\delta}_1 \in (0, \pi/m)$ for which all three of these phase velocities are identical: $c_1(\bar{\delta}_1) = c_2(\bar{\delta}_1) = c_3(\bar{\delta}_1)$. This $\bar{\delta}_1$ would correspond to a phase locked solution with maximal isotropy $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dotplus Z_m \dotplus Z_2^r$. A calculation shows that

$$c_1(\bar{\delta}_1) \;=\; k_1 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m} + 2\bar{\delta}_1\right) + k_2 \sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m} + \bar{\delta}_1\right) \tag{4.18}$$

$$c_2(\bar{\delta}_1) \;=\; 0 \tag{4.19}$$

$$c_3(\bar{\delta}_1) \;=\; -c_1(\bar{\delta}_1) \;. \tag{4.20}$$

Therefore, the desired solution exists if and only if $c_1(\bar{\delta}_1) = 0$ for some $\bar{\delta}_1 \in (0, \frac{\pi}{m})$. Such a $\bar{\delta}_1$ is guaranteed to exist if $c_1'(0)$ and $c_1'(\frac{\pi}{m})$ have the same sign. This condition on the derivatives of $c_1(\cdot)$ reduces to precisely the same requirement as that for λ_L and λ_R to take same sign in Proposition 4.3.3.

Example 1. For a fixed number m of blocks, let $f(\cdot) = \sin(m\cdot) + \sin(q\cdot)$, where q is not a multiple of m (in other words, $q \neq \mathcal{M}(m)$). Then,

$$\lambda_L \;=\; -m^2(2k_1 + k_2) \tag{4.21}$$

$$\lambda_R \;=\; -m^2(2k_1 - k_2) \;. \tag{4.22}$$

Therefore, if $k_1 > k_2/2$, there exists a solution of maximal isotropy $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dotplus Z_m(\rho_m) \dotplus Z_2^r(\rho_2^r)$. We verify this by choosing $m = 3$, $q = 2$, $k_2 = 3$, and $k_1 = 2 > k_2/2$ and plotting the quantity $c_1(\delta_1)$ discussed in Remark 4.3.1. See Figure 4.6(a); note that $c_1(\bar{\delta}_1) = 0$ for some $\bar{\delta}_1 \in (0, \pi/m)$, as expected.

Now, if we redistribute the phases among subblocks so that $k_2 = 5$, and $k_1 = 1 < k_2/2$, such a solution is no longer guaranteed; indeed, plotting $c_1(\delta_1)$ in this case (Figure 4.6(b)), we see that $c_1(\delta_1) \neq 0$ for $\delta_1 \in (0, \pi/m)$, so there is no solution with the desired maximal isotropy.

Example 2. For a fixed number m of blocks, let $f(\cdot) = \sin(2m\cdot) + \sin(q\cdot)$, where q is not a multiple of m. Then,

$$\lambda_L \;=\; -2m^2(2k_1 + k_2) \tag{4.23}$$

$$\lambda_R \;=\; -2m^2(2k_1 + k_2) \;. \tag{4.24}$$

Since λ_L and λ_R have the same sign for any k_1 and k_2, there will always exist a solution of maximal isotropy $(S_{k_1} \times S_{k_1} \times S_{k_2})^m \dotplus Z_m(\rho_m) \dotplus Z_2^r(\rho_2^r)$. We verify this via the plot of Figure 4.6(c) for $m = 3$, $q = 2$, $k_2 = 5$, and $k_1 = 1$: there exists a $\bar{\delta}_1 \in (0, \pi/m)$ such that $c_1(\bar{\delta}_1) = 0$, giving the expected phase locked solution.

Gradient dynamics and consequences

We observe as in Theorem 9.15 of [94] that:

Proposition 4.3.2. *Equation* (4.8) *is a gradient dynamical system on* T^N *with potential*

$$V = \frac{\alpha}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} F(\phi_j - \phi_i), \text{ where } f(\theta) = F'(\theta). \qquad (4.25)$$

Proof. Note that

$$-\frac{\partial V}{\partial \phi_i} = \frac{\alpha}{N} \sum_{j<i} f(\phi_j - \phi_i) - \frac{\alpha}{N} \sum_{j>i} f(\phi_i - \phi_j); \qquad (4.26)$$

the oddness of f implies that $\dot{\phi}_i = -\partial V / \partial \phi_i$. $\qquad \square$

Proposition 4.3.2 implies that

$$\dot{V} = \sum_{i=1}^{N} \frac{\partial V}{\partial \phi_i} \, \dot{\phi}_i = -\sum_{i=1}^{N} \dot{\phi}_i^2 \leq 0 \,, \qquad (4.27)$$

with equality only at equilibria. Thus, equation (4.8) with odd $f(\cdot)$ has no periodic or homoclinic orbits or heteroclinic cycles: all solutions approach equilibria, and almost all approach stable equilibria. In particular,

Corollary 4.3.3. *For f odd, equation* (4.7) *has no solutions unless $c = 0$.*

This immediately implies a corollary to Proposition 4.3.1, which shows that, for odd f, all phase locked solutions (not just fixed points) can be ruled out in the cylinder of phase space introduced there:

Corollary 4.3.4. *Let $R_1 > 0$ be such that* either *f restricted to $(0, 2R_1)$ or f restricted to $(-2R_1, 0)$ is of one sign (i.e., f is strictly negative or positive in the region). If f is odd* then there are no phase locked solutions of (4.8) in $\mathcal{C}_{R_1} \setminus \mathcal{D}$.

4.3.5 Linear stability of periodic orbits for pairwise phase-difference coupling

Rotating blocks

We study solutions with isotropy $(S_k)^m \dot{+} Z_m(\rho_m)$ (rotating block modes), $S_p \times S_{N-p}$ (two-block modes), and S_N (in-phase mode) for pairwise phase-difference systems (4.8); see Fig. 4.2. Since the Jacobian of (4.8) is constant along these periodic orbits with diagonal flow, this problem reduces to computation of eigenvalues ([6]). Stability will be discussed in terms of orbital stability, which implies asymptotic stability with respect to all perturbations transverse to the (continuous) T^1 group orbit of the

solution (hence excluding the corresponding zero eigenvalue). Note that if $c \neq 0$ in equation (4.7), the group orbit and periodic orbit coincide.

For $(S_k)^m \dot{+} Z_m$ symmetric solutions we have, as in [131] and (for $m = 1$) [174]:

Proposition 4.3.5. *Let $N = mk$ and let $\bar{\phi}$ be an $(S_k)^m \dot{+} Z_m(\rho_m)$-invariant fixed point or periodic orbit with diagonal flow. Then the eigenvalues of the Jacobian $J(\bar{\phi})$ obtained by linearization of Equation (4.8) are*

$$
\left.
\begin{array}{ll}
\lambda = \lambda_0 = 0, & \text{with multiplicity 1} \\
\lambda = \lambda_r^j, & j = 1, ..., m-1 \text{ : 'rotation eigenvalues'} \\
\lambda = \lambda_p, & \text{with multiplicity } m(k-1) \text{ : 'permutation eigenvalues'}
\end{array}
\right\},
$$

$$
\lambda_r^j = \frac{\alpha}{m} \sum_{k=1}^{m-1} f'\left(\frac{2\pi k}{m}\right)\left(\exp\left(\frac{2\pi k j i}{m}\right) - 1\right)
$$

$$
= \frac{\alpha}{2}\left(\sum_{l \in \mathcal{M}(m)_1^j} l\,(b_l^o + i b_l^e) + \sum_{l \in \mathcal{M}(m)_2^j} l\,(b_l^o - i b_l^e) - 2 \sum_{l \in \mathcal{M}(m)} l b_l^o\right) \tag{4.28}
$$

$$
\lambda_p = -\frac{\alpha}{m} \sum_{k=0}^{m-1} f'\left(\frac{2\pi k}{m}\right) = -\alpha \sum_{l \in \mathcal{M}(m)} b_l^o l, \tag{4.29}
$$

where

$$
\mathcal{M}(m)_1^j = \{mh - j | h = 1, 2, ...\}, \qquad \mathcal{M}(m)_2^j = \{mh + j | h = 0, 1, 2, ...\},
$$
$$
\mathcal{M}(m) = \{mh | h = 1, 2, ...\} .
$$

Remark 4.3.6. *A fact useful in interpreting Eqn. (4.28) is that unless m is even and $j = m/2$, conditions (ii) and (iii) above are not simultaneously violated by any l. Thus, unless $j = m/2$ the sets $\mathcal{M}_1^j$, $\mathcal{M}_2^j$, and $\mathcal{M}$ are mutually disjoint; for $j = m/2$, $\mathcal{M}_1^j \bigcap \mathcal{M}_2^j = \{m/2 + mh | h = 0, 1, 2, ...\}$.*

Remark 4.3.7. *As required since $J(\bar{\phi})$ is real, and as noted in [174], the rotational eigenvalues λ_j^r (excepting $\lambda_{N/2}^r$ if N is even) come in complex conjugate pairs. The relationship $\lambda_j^r = \overline{\lambda_{m-j}^r}$ follows from the facts that $l \in \mathcal{M}_2^j$ implies $l \in \mathcal{M}_1^{m-j}$ and $l \in \mathcal{M}_1^j$ implies $l \in \mathcal{M}_2^{m-j}$.*

The 'rotation' and 'permutation' terminology is due to [6], where general formulae for eigenvalues and eigenvectors are presented. The proof of Proposition 4.3.5 repeatedly uses the following simple fact. Define the set $\mathcal{M} = \{l | l = qm \text{ for some } q \in \mathbb{Z}\}$ and let $\gamma = \exp(2\pi i/m)$; then for $\bar{l} \in \mathbb{Z} \setminus \mathcal{M}$, $\sum_{r=1}^{m-1} \gamma^{\bar{l}r} = -1$. From here, the $m \times m$-blocked structure of the Jacobian along with results on the eigenvectors of Toeplitz matrices leads to the desired conclusion. For details, see the appendix to this chapter.

If $m = 1$, then the proposition addresses the S_N-invariant (in-phase) solutions. Here there are no rotation eigenvalues, and the permutation eigenvalues are simply

$$\lambda_p = -\alpha f'(0) , \tag{4.30}$$

with multiplicity $N - 1$. *Nonlinear* stability of these solutions is discussed in Section 4.4.2 (Proposition 4.4.2 with $\beta = 0$). At the other extreme, if $m = N$, the proposition addresses the Z_N-invariant solutions in which the phases of the oscillators are equally spaced; these are called 'rotating wave' solutions by [6], and correspond to the 'splay state' in the Josephson junction literature. In this case, there are no permutation eigenvalues, and (4.28) reduces to equation (63) of [174].

We now give examples to illustrate several interesting stability behaviors implied by Proposition 4.3.5. Including only the first harmonic in the coupling function ($f(\cdot) = b_1^o \sin(\cdot) + b_1^e \cos(\cdot)$), we have for $m > 1$:

$$\lambda_p = 0, \text{with multiplicity } m(k-1)$$

$$\lambda_r^j = \begin{cases} \frac{\alpha}{2}(b_1^o - ib_1^e) \text{ and } \frac{\alpha}{2}(b_1^o + ib_1^e) & \text{for } j = 1 \text{ and } m-1 \\ 0 & \text{otherwise (multiplicity } m-3), \end{cases}$$

in addition to λ_0. In this case the $(S_k)^m \dot{+} Z_m$ solutions are *highly degenerate* and, for $\alpha b_1^o > 0$, *unstable*. For $m = N$ ($k = 0$), there are $N - 2$ zero eigenvalues. This result is well-known from the Josephson junction literature; in [173], it is shown to be related to the integrability of the equations for this choice of f.

On the other hand, we note that inclusion of higher harmonics in $f(\cdot)$ generically unfolds the degeneracy in the sense that all but one (λ_0) of the eigenvalues become nonzero, implying instability or orbital stability. For example, adding the mth harmonic ($f(\cdot) = b_1^o \sin(\cdot) + b_1^e \cos(\cdot) + b_m^o \sin(m\cdot) + b_m^e \cos(m\cdot); \; m \neq 1$), we obtain

$$\lambda_p = -\alpha b_m^o m, \text{ with multiplicity } m(k-1)$$

$$\lambda_r^j = \begin{cases} \alpha[\frac{1}{2}(b_1^o - ib_1^e) - b_m^o m], \; c.c. & \text{if } j = 1, \; m-1 \\ -\alpha b_m^o m & \text{otherwise (multiplicity } m-3), \end{cases}$$

so that any $(S_k)^m \dot{+} Z_m$ solution is *orbitally stable* if $\alpha b_m^o > \alpha \frac{b_1^o}{2m}$ and $\alpha b_m^o > 0$. As a third example, consider the case where only odd Fourier modes are included in f ($b_l = 0$ for even l) and m is even. Then $\mathcal{M}(m)$ and $\mathcal{M}(m)_{1,2}^j$ contain only even integers for j even, and so the corresponding terms in (4.28)-(4.29) make no contribution. Thus, Proposition 4.3.5 shows that at least $m/2 - 1$ of the rotational eigenvalues are zero, and all $m(k-1)$ of the permutation eigenvalues are zero. With $\lambda_0 = 0$, this gives at least $N - m/2$ zero eigenvalues, so this is another degenerate case.

Remark 4.3.8. *For coupling functions whose harmonic indices belong entirely to $\mathcal{M}(l)$, any oscillator may be individually translated from a $(S_k)^m \dot{+} Z_m$ solution by a multiple of $\frac{2\pi}{l}$ to give another equilibrium. These translations give a total of l^N fixed points, each with identical stability (due to the $\frac{2\pi}{l}$ periodicity of f). Using calculations*

similar to those above, the eigenvalues are:

$$\lambda_p = 0, \, \text{with multiplicity } m(k-1)$$

$$\lambda_r^j = \left\{ \begin{array}{ll} \frac{\alpha l}{2}(b_l^o - ib_l^e), \; c.c. & \text{if } j = 1, \, m-1 \\ 0 & \text{otherwise (multiplicity } m-3) \end{array} \right\} .$$

We also note that the calculations which yield Proposition 4.3.5 also show which eigenvectors correspond to zero eigenvalues and hence along which directions there *may* be continuous families of equilibria. For example, with $k = 1$, $m = N = 4$ and $f(\cdot) = \sin(\cdot)$, the nondiagonal zero eigenvector is $(1, -1, 1, -1)^{\mathrm{T}}$, which reflects the fact that equilibrium is preserved if 'diametrically-opposite' pairs of oscillators are rotated independently.

Two-block periodic orbits

For $f'(0) \neq 0$, equation (4.30) guarantees that the S_N-invariant solutions satisfy the nondegeneracy assumption of Theorem 4.3.2. Then, the Theorem (with $m = 1$) implies that for some $\delta(p) > 0$, equation (4.8) has periodic orbits with $\phi_{ji} \in \{0, \, \delta(p), \, 2\pi - \delta(p)\}$ for all i, j. This occurs when two blocks of p and $N - p$ identical-phase oscillators are mutually out of phase by δ; to avoid redundancy, we restrict $0 \leq p \leq \lfloor N/2 \rfloor$. The Jacobian from linearizing around a $S_{N-p} \times S_p$ solution has a four-blocked structure which yields (see Appendix for details):

Proposition 4.3.9. *([103]) Let $\bar{\phi}$ be an $(S_p \times S_{N-p})$-invariant solution and $0 \leq p \leq \lfloor N/2 \rfloor$. Then for $p \geq 1$ the eigenvalues of the Jacobian from equation (4.8) are:*

$$\left. \begin{array}{lll} \lambda_1 = \alpha(b - \frac{p}{N}(a+b)), & \text{with multiplicity} & p-1 \\ \lambda_2 = \alpha(\frac{p}{N}(a+c) - a), & \text{with multiplicity} & N-p-1 \\ \lambda_3 = 0, & \text{with multiplicity} & 1 \\ \lambda_4 = \alpha(\frac{N-p}{N}b + \frac{p}{N}c), & \text{with multiplicity} & 1 \end{array} \right\} . \qquad (4.31)$$

Here, $a = f'(0)$, $b = -f'(\delta(p))$, $c = -f'(-\delta(p))$ *We remark that the result for $p = 0$ also follows from Proposition 4.3.5.*

If $f(\cdot)$ is odd, two-block states with $\delta = \pi$ exist for any p since $f(0) = f(\pi) = 0$; we write $\delta \neq \delta(p)$ to indicate this p-independence of δ. Oddness of f also implies $b = c$. This case was studied in [131], where expressions corresponding to (4.31) are presented.

Corollary 4.3.10. *Assume that $b = c$, $\delta \neq \delta(p)$, and that $a, b > 0$. If $\alpha > 0$, the two-block equilibria of equation (4.8) are orbitally stable if and only if $p = 0$. If $\alpha < 0$, the equilibria are stable if and only if $p \neq 0$ and $a < bp/(N-p)$, if the equilibria are stable for $p = k$ for some $k \leq \lfloor N/2 \rfloor$, then they are stable for $p > k$.*

Proof. The results for $\alpha > 0$ are immediate from λ_4 of equation (4.31) and (4.30). For $\alpha < 0$, we note that $\lambda_{1,2} \leq 0$ implies $Na \leq p(a+b) \leq Nb$. Upon rearranging, this yields $a \leq b(N-p)/p$ and $a \leq bp/(N-p)$; for p in the given range, the latter

inequality implies the first, and for fixed a and b it is clear that if the second inequality is satisfied for $p = k$, then it continues to be satisfied as p increases. In this case λ_1, λ_2 and λ_4 are all strictly negative, leading to the Corollary. $\square$

We remark that if $a, b < 0$, the sign of α may be switched and the Corollary applied, and that the result that stability of equilibria for $p = k$ implies stability for $p = N/2$ is stated in [131].

The corollary indicates that for $\alpha < 0$ and under certain conditions on a, b, and N, orbital stability of two-block fixed points can change as p is varied. For example, if $a = 1$, $b = 2$, and $N = 5$, the equilibria are unstable for $p = 0, 1$ but are stable for $p = 2$. In the special case $a = b = c$ (which occurs, for example, if $f(\cdot) = \sin(\cdot)$), note that $\lambda_1 = -\lambda_2 = \alpha(a - 2p/N)$, $\lambda_4 = \alpha a$; thus the fixed points are unstable unless $\alpha a < 0$, N is even and $p = N/2$, in which case they are neutrally stable with $N - 1$ zero eigenvalues. As above, inclusion of higher harmonics in the Fourier series for $f(\cdot)$ generically unfolds this degeneracy.

We close this subsection by remarking that techniques used to prove Propositions 4.3.5 and 4.3.9 could in principle be extended to calculate the stability of general $(S_{k_1} \times S_{k_2})^m \dotplus Z_m$ solutions for $m > 1$, where $m(k_1 + k_2) = N$. We refer the reader to [6] for the specific example $(S_2 \times S_1)^3 \dotplus Z_3$.

4.3.6 Nonlinear stability of periodic orbits – domains of attraction

Global stability for $\sin$ **coupling**

In the special case that $f(\cdot) = \sin(\cdot)$, it is possible to characterize the global dynamics of the system (4.8). In particular, Watanabe and Strogatz [173] define the (codimension 2) incoherent manifold as the set where the centroid of the phases is zero:

$$r \equiv \sum_j \exp(i\phi_j) = 0 \tag{4.32}$$

$(i = \sqrt{-1})$ and the (codimension $N-1$) coherent manifold as the set where $\phi_j \equiv c$, $\forall j$, for some c. We will also refer to the coherent manifold as the perfectly synchronized state (i.e., the diagonal $\mathcal{D}$ in Fig. 4.3). Watanabe and Strogatz show, using a change of coordinates, that almost every initial condition asymptotically approaches the incoherent manifold when $\alpha < 0$ (i.e., the incoherent manifold is globally attracting in this case), and that the coherent manifold is globally attracting when $\alpha > 0$. Here, we obtain these results via an elementary proof independent of the Watanabe-Strogatz coordinate change. The present method of proof, based on the gradient function of Prop. 4.3.2 (valid since $\sin(\cdot)$ is odd), is a special case of that independently used in [98] to obtain more general results. We show

Proposition 4.3.4. *The incoherent manifold is globally attracting when $\alpha < 0$.*

Proposition 4.3.5. *The coherent manifold $\mathcal{D}$ (i.e. perfectly synchronized state) is globally attracting when $\alpha > 0$.*

Proposition 4.3.4 will find application in Section 5.4.2 below.

First, we classify the equilibria of

$$\dot{\phi}_k = \frac{\alpha}{N} \sum_{j=1}^{N} \sin\left(\phi_j - \phi_k\right) \ . \tag{4.33}$$

Note that the condition $\dot{\phi}_k = 0 \ \forall \ k$ implies

$$\left[\sum_j e^{i(\phi_j - \phi_k)} - \sum_j e^{i(\phi_k - \phi_j)} \right] = 0 \tag{4.34}$$

$$\sum_j e^{i(\phi_j - \phi_k)} = \sum_j e^{i(\phi_k - \phi_j)} \tag{4.35}$$

$$e^{-i\phi_k} \sum_j e^{i\phi_j} = e^{i\phi_k} \sum_j e^{-i\phi_j} \tag{4.36}$$

This final equation is clearly satisfied for all states on the incoherent manifold (in which case $\sum_j \exp(i\phi_j) = 0$ which immediately implies $\sum_j \exp(-i\phi_j) = 0$). That is, the incoherent $N-2$ dimensional incoherent manifold consists entirely of fixed points.

For equilibria not on the incoherent manifold, i.e. $\sum_j \exp(i\phi_j) \neq 0$, (4.36) gives

$$e^{-2i\phi_k} = \frac{\sum_j e^{-i\phi_j}}{\sum_j e^{i\phi_j}} \quad \forall k \ . \tag{4.37}$$

Note that the right hand side is independent of k. This implies that $\phi_k = c + n_k \pi$ $\forall \ k$, for some integers n_k. In other words, all phases differ by 0 or π. Hence, the only equilibria not on the incoherent manifold are "two-block" states or are on the synchronized manifold (when all phases differ by 0).

In summary, we have shown

Lemma 4.3.1. *For $\alpha \neq 0$, all equilibria of Eqn. (4.33) are either (i) on the incoherent manifold, or (ii) are "two-block states" with $\phi_1, ..., \phi_p \equiv c$ and $\phi_{p+1}, ..., \phi_N \equiv c + \pi$ for some $c \in [0, 2\pi)$ and some value of $p \in \{1, ..., \lceil N/2 \rceil - 1\}$, or (iii) are on the coherent manifold.*

Note that the incoherent manifold contains states with no permutation or cyclic symmetries; i.e., solutions other than the rotating-block or two-block types (for example, $N = 5$ and $\theta_1 = 0$, $\theta_2 = \theta_3 = \delta$, $\theta_4 = \theta_5 = -\delta$, where $\delta = \cos^{-1}\left(\frac{1}{4}\right)$ and $\delta < \pi$).

Proof of Proposition 4.3.4. A direct calculation (see Corollary 4.3.10 above) shows that, for the system (4.33), the coherent manifold and all two block states are unstable unless N is even and each of the two blocks (separated by π) contains the same number $N/2$ of phases. These latter states are on the incoherent manifold. Together with Lemma 4.3.1, this shows that all equilibria of (4.33) not on the incoherent manifold are unstable, so that all stable equilibria lie in the incoherent manifold.

Thus, since almost every initial condition flowing under the gradient dynamics of (4.33) must a approach a stable equilibrium, we have our desired result: the incoherent manifold is globally attracting. □

Proof of Proposition 4.3.5. To obtain this result we must show that all two-block states ($p > 0$) and all states on the incoherent manifold are unstable. Therefore almost all initial conditions must flow to the only possible stable equilibria, the coherent manifold. As above, direct calculation [22] shows that the two-block states are unstable with $\alpha > 0$. We now show the instability of all incoherent states.

It suffices to show that the potential V obtains its global maximum value only on the incoherent manifold. Then, if a trajectory initially on the incoherent manifold is perturbed off of this manifold, the potential will decrease, and, due to (4.27), the trajectory will never return to the incoherent manifold, demonstrating its instability.

This property of V is easy to show. Since it is defined over the torus, the extreme values of V occur (only) where $\nabla V = 0$, i.e., at equilibria of (4.33). We now calculate V at these equilibria.

$$V = -\frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k \geq j} \cos(\phi_j - \phi_k) \tag{4.38}$$

$$= -\frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \cos(\phi_j - \phi_k) + \frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k<j} \cos(\phi_j - \phi_k) \tag{4.39}$$

$$= -\frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \cos(\phi_j - \phi_k) + \frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k \leq j} \cos(\phi_j - \phi_k) - \alpha \tag{4.40}$$

$$= -\frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \cos(\phi_j - \phi_k) + \frac{\alpha}{N} \sum_{j=1}^{N} \sum_{k \geq j} \cos(\phi_j - \phi_k) - \alpha \tag{4.41}$$

so that

$$V = -\frac{\alpha}{2N} \sum_{j=1}^{N} \sum_{k=1}^{N} \cos(\phi_j - \phi_k) - \frac{\alpha}{2} \tag{4.42}$$

We now compute V at one of the two-block equilibria (groups of p and $N-p$ angles out of phase by π) described in Lemma 4.3.1. To directly calculate the sum in Eqn. (4.42) we note that, each of the p times that ϕ_j lies in the block of p phases, the argument of the sum over k takes the value 1, p times, and it takes the value -1, $N - p$ times, (and similarly when ϕ_j lies in the block of $N - p$ phases). Thus, at a two-block state,

$$V = -\frac{\alpha}{2N} [p(p - (N - p)) + (N - p)((N - p) - p)] - \frac{\alpha}{2} \tag{4.43}$$

$$= -\frac{\alpha}{2N} [p - (N - p)]^2 - \frac{\alpha N}{2N} \leq -\frac{\alpha}{2} , \tag{4.44}$$

with equality only if N is even and $p = N/2$ (in which case the two-block state lies

on the incoherent manifold). Note that plugging $p = 0$ into (4.44) gives the potential on the coherent manifold.

On the incoherent manifold, we expand the cos terms in (4.42), giving

$$
\begin{aligned}
V &= -\frac{\alpha}{2N} \sum_{j=1}^{N} \cos(\phi_j) \sum_{k=1}^{N} \cos(\phi_k) + \frac{\alpha}{2N} \sum_{j=1}^{N} \sin(\phi_j) \sum_{k=1}^{N} \sin(\phi_k) - \frac{\alpha}{2} \quad (4.45) \\
&= 0 + 0 - \frac{\alpha}{2} \; . \quad (4.46)
\end{aligned}
$$

Thus, comparing (4.44) and (4.46), which give all of the (local and global) extrema of V, we see, as desired, that the potential reaches its maximum, $-\alpha/2$, only on the incoherent manifold and everywhere on this manifold, so the proposition follows. $\square$

Estimate for the domain of attraction of $\mathcal{D}$ for arbitrary phase-difference coupling functions f

We now give a nonlinear stability result for the globally synchronized state that allows for more general coupling. The following is valid for systems of the form (4.8):

Proposition 4.3.11. *(Nonlinear stability of $\mathcal{D}$.) For some $s > 1$, assume $f'(0) > 0$ and let $R > 0$ be the smallest value for which either $f'(2R) = f'(0)/s > 0$ or $f'(-2R) = f'(0)/s > 0$ (implying $\min_{\theta \in [-2R, 2R]} f'(\theta) = f'(0)/s$). Then, if*

$$
\alpha > 0
$$

the domain of attraction for $\mathcal{D}$ includes $\mathcal{C}_R \overset{\triangle}{=} \{\theta | d(\theta, \theta_d(\psi)) \leq R \text{ for some } \psi \in [0, 2\pi]\}$ (cf. Figure 4.3).

Proposition 4.3.11 follows from setting $\beta = 0$ in the more general Proposition 4.4.2, the proof of which is given below.

4.3.7   Existence of fixed l_B-tori

Proposition 4.3.12. *For ϕ contained in an invariant l_B-torus $\mathrm{Fix}[(S_{k_1} \times \cdots \times S_{k_{l_B}})^m \dot{+} Z_m]$ with $N = m(k_1 + \cdots + k_{l_B})$, equation (4.8) reduces to:*

$$
\dot{\phi}_i = \frac{\alpha}{N} \sum_{l \in \mathcal{M}(m)} \left\{ b_l^e m \sum_{q=1}^{l_B} k_q(i) \cos[l x_q(i)] + b_l^o m \sum_{q=1}^{l_B} k_q(i) \sin[l x_q(i)] \right\}, \quad (4.47)
$$

where the numbers $k_q(i)$ and the angles $x_q(i)$ are as explained in Figure 4.7. In particular (as found in [6]), if $b_l^{e,o} = 0$ for all $l \in \mathcal{M}(m)$, then the l_B-torus is a continuum of fixed points.

Proof. Consider $\dot{\phi}_i = \alpha \sum_{j=1}^{N} f(\phi_{ji})$ for arbitrary i. First we calculate the contributions to $\dot{\phi}_i$ from odd (sine) modes of the coupling function. Computing the vectorfield

70
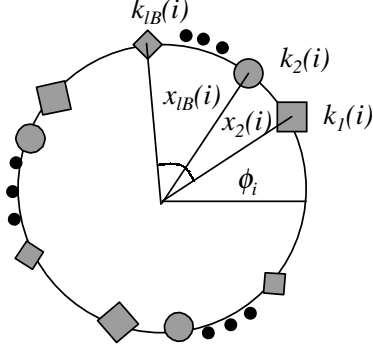
Figure 4.7: The labeling scheme used in Proposition 4.3.12. Given the reference index i corresponding to the $\dot{\phi}_i$ being computed, blocks of oscillators are numbered by the index q in a counterclockwise fashion, starting with $q = 1$ for the block containing ϕ_i itself. Each block contains $k_q(i)$ oscillators and is separated from its neighbor by the angle $x_q(i)$ (by definition $x_1(i) \equiv 0$).

in Eqn. (4.8) separately for each cluster of m blocks of $k_q(i)$ oscillators, the lth Fourier mode adds to $\dot{\phi}_i/[\alpha k_q(i)]$:

$$
b_l^o \sum_{j=0}^{m-1} \sin\left(lx_q(i) + l\frac{2\pi j}{m}\right) = b_l^o \sin(lx_q(i)) \sum_{j=0}^{m-1} \cos\left(l\frac{2\pi j}{m}\right)
$$
$$
+ \; b_l^o \cos(lx_q(i)) \sum_{j=0}^{m-1} \sin\left(l\frac{2\pi j}{m}\right)
$$
$$
= \frac{b_l^o}{2} \sin(lx_q(i)) \sum_{j=0}^{m-1} \left(\gamma^{lj} + \gamma^{-lj}\right) \; + \; \frac{b_l^o}{2i} \cos(lx_q(i)) \sum_{j=0}^{m-1} \left(\gamma^{lj} - \gamma^{-lj}\right) , \quad (4.48)
$$

where $\gamma = e^{2\pi i/m}$. But, from Lemma 4.5.1, $\sum_{j=0}^{m-1} \gamma^{lj} = 0$ if $l \notin \mathcal{N}(m)$ and $= m$ otherwise. Summing over the l_b clusters, this gives the contribution of the odd Fourier modes in Eqn. (4.47).

Repeating this analysis for the even modes, we obtain

$$
b_l^e \sum_{j=0}^{m-1} \cos\left(lx_q(i) + l\frac{2\pi j}{m}\right) = b_l^e \cos(lx_q(i)) \sum_{j=0}^{m-1} \cos\left(l\frac{2\pi j}{m}\right)
$$
$$
- \; b_l^e \sin(lx_q(i)) \sum_{j=0}^{m-1} \sin\left(l\frac{2\pi j}{m}\right)
$$
$$
= \frac{b_l^e}{2} \cos(lx_q(i)) \sum_{j=0}^{m-1} \left(\gamma^{lj} + \gamma^{-lj}\right) \; + \; \frac{b_l^e}{2i} \cos(lx_q(i)) \sum_{j=0}^{m-1} \left(\gamma^{lj} - \gamma^{-lj}\right) . \quad (4.49)
$$

By the same remarks following Eqn. (4.48), only the first term in Eqn. (4.49) contributes to the flow, and only when $l \in \mathcal{N}(m)$. This yields the remainder of Eqn. (4.47).

Because the x_q are arbitrary spacings between l_B blocks, the family of fixed points considered in the proof is at least $l_B - 1$-dimensional. But the 'initial' ϕ_i in Fix$[\Sigma_{\mathbf{k},m}]$ may also take arbitrary values in Fix$[\Sigma_{\mathbf{k},m}]$, so the entire set is a l_B-torus. $\square$

For odd f, the fixed tori may also be found by showing that the potential (4.25) is always constant under this same condition on the Fourier coefficients of f given in Proposition 4.3.12.

If $m = N$ and $b_l^{e,o} = 0$ for all $l = 0$ (mod N), $l_B = 1$ and Proposition 4.3.12 simply gives the circle of equilibria that is the T^1 group orbit of the Z_N-symmetric equilibrium of Proposition 4.3.5 (with $k = 1$). If $m = 1$ then Proposition 4.3.12 gives no new information about fixed subspaces: $b_l^{e,o} = 0$ for all $l \in \mathcal{M}(m)$ implies that the oscillators are uncoupled.

It is a general fact that if $G_1 \subseteq G_2$ for isotropy subgroups G_1 and G_2, then Fix$(G_2) \subseteq$ Fix(G_1). Therefore, since $S_1 \times S_1 \times \cdots \times S_1 \subseteq S_{k_1} \times \cdots \times S_{k_{l_B}}$, the l_B-tori of fixed points guaranteed by the theorem are actually contained in the (N/m)-torus Fix$[(S_1 \times S_1 \times \cdots \times S_1)^m \dot{+} Z_m] =$ Fix$[Z_m]$.

The following examples illustrate implications of Proposition 4.3.12.

Example 1. Consider $N = 4$ and suppose $b_l^{e,o} = 0$ for even l. For $m = 4$ the torus of fixed points guaranteed by the proposition is just the one-torus Fix$[Z_4]$. For $m = 2$, we get the two-torus of fixed points Fix$[Z_2]$. This describes the set of points for which two oscillators are out of phase by π, and the other two are also out of phase by π, corresponding to $(\phi_1, \phi_2, \phi_3, \phi_4) = (\xi_1, \xi_2, \xi_1 + \pi, \xi_2 + \pi)$. Fix$[Z_2]$ contains both Fix$[Z_4] = \{(\xi, \xi + \pi/2, \xi + \pi, \xi + 3\pi/2)\}$ and Fix$[(S_2)^2 \dot{+} Z_2] = \{(\xi, \xi, \xi + \pi, \xi + \pi)\}$.

Fix$[Z_2]$ also coincides with the $(N - 2 = 2)$-dimensional incoherent manifold described in Section 4.3.6. Because Fix$[Z_2]$ is a fixed point subspace, the two-dimensional incoherent manifold is dynamically invariant as found in [174]; Proposition 4.3.12 gives conditions under which it is also dynamically fixed as well as the expression for drift along this manifold. [174] also show that the $(N - 2)$ dimensional incoherent manifold is not dynamically invariant when $N \geq 5$.

However, this manifold contains dynamically invariant (and perhaps dynamically fixed) submanifolds: for ϕ in fixed point subspaces of isotropy subgroups which have Z_m as a subgroup (where $m \geq 2$), the relevant centroid is zero. Thus, these fixed point subspaces are contained in the incoherent manifold. Note that the invariant (or fixed) tori have dimension $l_B \leq N/m$, which is less than $N - 2$ for $N \geq 5$, $m \geq 2$.

Example 2. Suppose $N = 6$ and $f(\cdot) = \sin(\cdot)$, and consider the $((S_3)^2 \dot{+} Z_2)$-invariant equilibria (e.g., $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (0, 0, 0, \pi, \pi, \pi) \equiv \bar{\phi}$). From Proposition 4.3.5, the eigenvalues for such equilibria are 0 with multiplicity five, and 6α with multiplicity one. The null eigenvectors may be taken to be $e_1 = (1, 1, 1, 1, 1, 1)$, $e_2 = (2, -1, -1, 0, 0, 0)$, $e_3 = (1, 1, -2, 0, 0, 0)$, $e_4 = (2, -1, -1, 2, -1, -1)$, $e_5 = (1, 1, -2, 1, 1, -2)$. Figure 4.8 shows the potential V corresponding to perturbations to $\bar{\phi}$ in the directions of these null eigenvectors. V is flat for perturbations in the e_1, e_4, and e_5 directions (each with a corresponding one-dimensional continuum of fixed points, overall giving
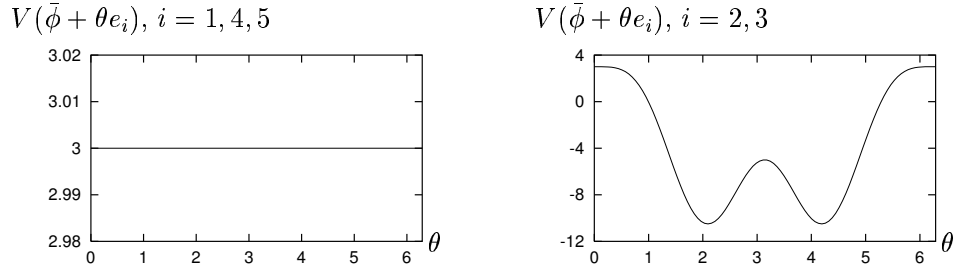
Figure 4.8: Potential V for perturbations to $\bar{\phi}$ in the directions of the null eigenvectors, as defined in the text.

a three-torus of equilibria), but not for perturbations in the e_2 and e_3 directions. Proposition 4.3.12 guarantees the existence of the three-torus of equilibria Fix$[Z_2]$ given by $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (\xi_1, \xi_2, \xi_3, \xi_1 + \pi, \xi_2 + \pi, \xi_3 + \pi)$; note that perturbations to $\bar{\phi}$ in the e_1, e_4, and e_5 directions keep the system in the Fix$[Z_2]$ subspace. The e_2 and e_3 perturbations illustrate that every zero eigenvalue of Propositions 4.3.5 or 4.3.9 does *not* necessarily imply a corresponding one-dimensional continuum of fixed points.

Example 3. Intersection of fixed tori. Suppose $N = 6$ and $b_l^{e,o} = 0$ for even l and also for $l = 0 \pmod 3$ (for example, one could take $f(\cdot) = b_1 \sin(\cdot) + b_5 \sin(5\cdot) + b_7 \sin(7\cdot)$). For $m = 2$, the proposition implies that there is a three-torus of fixed points Fix$[Z_2]$ as given in Example 2. For $m = 3$, the proposition implies that there is a two-torus of fixed points Fix$[Z_3]$ given by $(\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (\xi_4, \xi_5, \xi_4 + 2\pi/3, \xi_5 + 2\pi/3, \xi_4 + 4\pi/3, \xi_5 + 4\pi/3)$ and permutations. The intersection of Fix$[Z_2]$ and Fix$[Z_3]$ must at least include Fix$[Z_6]$ since $Z_2 \subset Z_6$, $Z_3 \subset Z_6 \Rightarrow$Fix$[Z_6] \subseteq$ Fix$[Z_2]$, Fix$[Z_6] \subseteq$ Fix$[Z_3]$.

4.4 Breaking the T^1 Symmetry: Product Coupling

Reintroducing the h and g terms and going back to θ coordinates, we return to the S_N-equivariant Eqn. (4.2), which we reproduce here for easy reference:

$$\dot{\theta}_i = \omega + \frac{\alpha}{N} \sum_{j=1}^{N} f(\theta_j - \theta_i) + h(\theta_i) \frac{\beta}{N-1} \sum_{j \neq i}^{N} g(\theta_j). \tag{4.50}$$

The results of this section are valid for arbitrary C^1 2π-periodic functions g and h; without loss of generality, we assume that g takes values in $[0, 1]$. Additional assumptions on the product function $G(\theta) \triangleq h(\theta)g(\theta)$ simplify the discussion of bifurcations in Section 4.4.1.

4.4.1 Bifurcations of fixed points on the diagonal $\text{Fix}[S_N]$

This section concerns analysis local to the (perfectly synchronized) diagonal of T^N, defined by $\mathcal{D} \triangleq \{\theta | \theta_i = \theta_j \; \forall \; i, j\}$, which is dynamically invariant (see Fig. 4.3). Restricted to $\mathcal{D}$ and with $\theta_i \equiv \theta$, Eqn. (4.50) becomes

$$\dot{\theta} = \omega + \alpha f(0) + \beta G(\theta). \tag{4.51}$$

This equation has fixed points given by $\bar{\theta} = G^{-1}\left(-\frac{\omega + \alpha f(0)}{\beta}\right)$. To simplify the analysis in this section, we assume that G has a single minimum θ_{min} with $G''(\theta_{min}) \neq 0$, as it does for the 'neurobiological' coupling functions to be considered in Section 4.5.

These conditions on G and $\omega > 0$ imply that there are no, one, or two on-diagonal fixed points, in the latter case denoted by $\bar{\theta}_1 < \bar{\theta}_2$. Linearized about these fixed points (when they exist) $\bar{\theta}_k$, Eqn. (4.50) is

$$\begin{aligned}
\dot{\xi}_i &= (N-1)\left[-\frac{\alpha}{N}f'(0) + \frac{\beta}{N-1}g(\theta_k)h'(\theta_k)\right]\xi_i + \\
&\quad \left[\frac{\alpha}{N}f'(0) + \frac{\beta}{N-1}h(\theta_k)g'(\theta_k)\right]\sum_{j \neq i}\xi_j \\
&\triangleq [A_d\xi]_i \;,
\end{aligned} \tag{4.52}$$

where

$$A_d = \begin{bmatrix} a_k & c_k & c_k & \cdots & c_k \\ c_k & a_k & c_k & \cdots & c_k \\ \cdots & & & & \\ c_k & c_k & c_k & \cdots & a_k \end{bmatrix}, \tag{4.53}$$

and

$$a_k = (N-1)\left[-\frac{\alpha}{N}f'(0) + \frac{\beta}{N-1}g(\theta_k)h'(\theta_k)\right] \quad, \quad c_k = \frac{\beta}{N-1}h(\theta_k)g'(\theta_k) + \frac{\alpha}{N}f'(0) \;. \tag{4.54}$$

Using the same argument as in the proof of Proposition 4.3.9, the eigenvalues of A_d are

$$\lambda_1^k = -\alpha f'(0) + \beta\left(h'(\bar{\theta}_k)g(\bar{\theta}_k) - \frac{1}{N-1}h(\bar{\theta}_k)g'(\bar{\theta}_k)\right), \text{multiplicity } N-1, \tag{4.55}$$

$$\lambda_2^k = \beta G'(\bar{\theta}_k), \text{multiplicity } 1. \tag{4.56}$$

Stability in the transverse directions (with respect to the diagonal) is determined by λ_1, and in the axial direction by λ_2; note that our hypothesis on G implies that $\lambda_2(\bar{\theta}_1) < 0$, and $\lambda_2(\bar{\theta}_2) > 0$. As β is decreased through $\beta = (\omega + \alpha f(0))/|G(\theta_{min})|$, the two fixed points coalesce and disappear in a saddle node bifurcation (the appearance of these fixed points as β increases represents the phenomenon of oscillator death: [53, 164]). For the remaining values of β, the orbit along $\mathcal{D}$ is the (S_N symmetric) periodic orbit $\theta_D(t)$. We will investigate the period and stability of this orbit in the

following sections.

4.4.2 Frequency and stability of the in-phase periodic orbit

If $\beta < (\omega + \alpha f(0))/|G(\theta_{min})|$, the period of the orbit on $\mathcal{D}$ is given by

$$\tau = \int_0^{2\pi} \left(\frac{d\theta}{dt} \right)^{-1} d\theta = \int_0^{2\pi} \frac{d\theta}{\omega + \alpha f(0) + \beta h(\theta) g(\theta)}. \tag{4.57}$$

Moreover, we have

Proposition 4.4.1. *(Local stability of $\mathcal{D}$.)* *The S_N-symmetric periodic solution* $\theta_i(t) \equiv \theta_D(t)$ *along $\mathcal{D}$ is asymptotically stable if*

$$\alpha > \frac{\beta N}{(N-1)\tau f'(0)} \int_0^{2\pi} \frac{g(\theta)h'(\theta)}{\omega + \alpha f(0) + \beta g(\theta) h(\theta)} d\theta, \tag{4.58}$$

where τ is the (generally α-dependent) period of $\theta_D(t)$ given in equation (4.57) and we assume $f'(0) > 0$.

Closely related results are found in [166, 75].

Proof. Linearized around $\theta_D(t)$, equation (4.50) becomes $\dot{\xi}_i = [A_d(t)\xi]_i$. The proof uses the fact that the (t-dependent) symmetric matrix $A_d(t)$ has a particularly simple structure, and that it can be diagonalized by a t-independent similarity transformation. Specifically, the eigenvalues of $A_d(t)$, where t is viewed as a (fixed) parameter, are:

$$\lambda_1(t) = -\alpha f'(0) + \beta \left(g(t)h'(t) - \frac{1}{N-1} h(t)g'(t) \right), \quad \text{multiplicity } N-1,$$

$$\lambda_2(t) = \beta G'(t), \quad \text{multiplicity } 1,$$

where $g(t)$ is written for $g(\theta_D(t))$, etc. The orthogonal eigenvectors of $\lambda_1(t)$ (denoted by $\chi_1, \ldots, \chi_{N-1}$) may be chosen constant and orthogonal to the eigenspace of $\lambda_2(t)$, which is spanned by the eigenvector $(1, \ldots, 1)^T$. Thus, $\chi_1, \ldots, \chi_{N-1}$ span the space normal to $\theta_D(t)$. In these eigencoordinates the linearized system decouples as

$$\dot{\xi}_i = \lambda_1(t)\xi_i, \ i = 1, \ldots, N-1, \ \ \dot{\xi}_N = \lambda_2(t)\xi_N.$$

Define the $(N-1)$-dimensional plane $\sum = \{\chi | \chi_N = 0\}$ and consider the Poincaré map $P : U \to U$ for some neighborhood $U \subset \sum$ of 0. The orbit $\theta_D(t)$ intersects $\sum$ at 0, which is a fixed point for P. For $i = 1, \ldots, N-1$, $P : \xi_i \mapsto (\exp \int_0^\tau \lambda_1(t)dt)\xi_i$ so that 0 is a stable fixed point for P if $\int_0^\tau \lambda_1(t)dt < 0$. (Due to the periodicity of $G(t)$, $\int_0^\tau \lambda_2(t)dt \equiv 0$, as it must, being the Floquet exponent along the periodic orbit). We

have

$$\int_0^\tau \lambda_1(t)dt$$

$$= \int_0^\tau \left(-\alpha f'(0) + \beta \left[g(t)h'(t) - \frac{1}{N-1} h(t)g'(t) \right] \right) dt$$

$$= -\alpha f'(0)\tau + \int_0^{2\pi} \beta \left(-\frac{[h(\theta)g'(\theta) + h'(\theta)g(\theta)]}{N-1} + \frac{Ng(\theta)h'(\theta)}{N-1} \right) \dot\theta^{-1} d\theta$$

$$= -\alpha f'(0)\tau - \frac{1}{N-1} \ln[\omega + \alpha f(0) + \beta h(\theta)g(\theta)]_0^{2\pi}$$

$$+ \int_0^{2\pi} \frac{\beta N g(\theta)h'(\theta)}{(N-1)(\omega + \alpha f(0) + \beta h(\theta)g(\theta))} d\theta \qquad (4.59)$$

$$= -\alpha f'(0)\tau + \int_0^{2\pi} \frac{\beta N g(\theta)h'(\theta)}{(N-1)(\omega + \alpha f(0) + \beta h(\theta)g(\theta))} d\theta , \qquad (4.60)$$

where the second term in equation (4.59) vanishes due to the 2π-periodicity of h and g. Thus, $\int_0^\tau \lambda_1(t)dt < 0$ when the inequality of Proposition 4.4.1 is satisfied. Since stability of the fixed point 0 under P implies stability of $\theta_D(t)$ for equation (4.8), the Proposition is proven. $\qquad\square$

A simple calculation using integration by parts and the 2π-periodicity of g and h shows that for $\alpha = 0$ and asymptotically small β, the right-hand side of (4.60) becomes $\frac{\beta N}{N-1} f'_s(0)$, which (cf. (4.30)) determines the stability of the in-phase solution if synaptic coupling $\frac{\beta}{N-1} h(\theta_i) \sum_{j\neq i} g(\theta_j)$ is taken to be weak and then averaged to yield $\frac{\beta}{N-1} \sum_{j\neq i} f_s(\theta_j - \theta_i)$. This agreement between the averaged and original versions of (4.50) for sufficiently small β is expected from the averaging theorem ([82]), and reveals how (4.58) generalizes the stability result found in [171, 85] for $N = 2$ and averaged synaptic coupling.

Equation (4.58) may be used to estimate a critical value α_{loc} such that $\theta_D(t)$ is asymptotically stable for $\alpha > \alpha_{loc}$. Letting $\widehat{h}$ be a Lipschitz constant for h, we note that

$$\int_0^{2\pi} \frac{g(\theta)h'(\theta)}{\omega + \alpha f(0) + \beta h(\theta)g(\theta)} d\theta \le \int_0^{2\pi} \frac{\widehat{h}}{\omega + \alpha f(0) + \beta h(\theta)g(\theta)} d\theta = \widehat{h}\tau, \qquad (4.61)$$

where the inequality follows from the bound on g and the definition of the Lipschitz constant. Thus, from (4.58) we have stability if $\alpha > \frac{N\beta\widehat{h}}{(N-1)f'(0)} \stackrel{\triangle}{=} \alpha_{loc}$. With $f(0) = 0$ (e.g. if f is odd), this estimate can be refined: the right-hand side of equation (4.58) is independent of α, so that the (smallest) critical value $\tilde\alpha_{loc}$ is

$$\tilde\alpha_{loc} = \frac{\beta N}{(N-1)\tau f'(0)} \int_0^{2\pi} \frac{g(\theta)h'(\theta)}{\omega + \beta g(\theta)h(\theta)} d\theta. \qquad (4.62)$$

We now turn to the nonlinear stability properties of $\mathcal{D}$.

Estimate for the domain of attraction of $\mathcal{D}$

Proposition 4.4.2. *(Nonlinear stability of $\mathcal{D}$.) For some $s > 1$, assume $f'(0) > 0$ and let $R > 0$ be the smallest value for which either $f'(2R) = f'(0)/s > 0$ or $f'(-2R) = f'(0)/s > 0$ (implying $\min_{\theta \in [-2R, 2R]} f'(\theta) = f'(0)/s$). Let $\widehat{G}$ be the Lipschitz constant for $G(\cdot) = g(\cdot)h(\cdot)$, and define $\widehat{h}_1(\theta_i) = \max_\theta\{|h'(\theta_i)g(\theta)| : |\theta_i - \theta| < 2R\}$ and $\widehat{h}_1 = \max_{\theta_i}\{\widehat{h}_1(\theta_i)\}$. Then, for*

$$\alpha > \alpha_{glob} \triangleq \frac{s\beta(N\widehat{h}_1 + \widehat{G})}{(N-1)f'(0)} \ , \tag{4.63}$$

the domain of attraction for $\mathcal{D}$ includes $\mathcal{C}_R \triangleq \{\theta | d(\theta, \theta_d(\psi)) \leq R \text{ for some } \psi \in [0, 2\pi]\}$ (cf. Figure 4.3).

Proof. Fix an arbitrary $\psi \in [0, 2\pi)$. Consider the (non-orthogonal) basis $b \equiv \{x_i | i = 1, \ldots, N-1\}$, where $x_i \equiv \theta_i - \theta_{i+1}$. We define X_ψ, the $N-1$ dimensional space perpendicular to the axis of $\mathcal{C}_R$ at $\theta_d(\psi)$, as the copy of span b containing $\theta_d(\psi)$. In other words, X_ψ is the normal space $\mathbf{N}(\theta_d(\psi))$.

Now, define the squared 'radius' $\mathcal{R} = \sum_{i=0}^{N-1} x_i^2$. We will show that $\dot{\mathcal{R}} = 2\sum_{i=0}^{N-1} x_i \dot{x}_i \leq 0$ for all $\mathbf{x} \in \mathcal{C}_R$. The cylindrical surfaces $\{\mathbf{x} | \mathcal{R}(\mathbf{x}) = c\}$ will therefore be crossed 'inward' toward the axis of $\mathcal{C}_R$.

Take an arbitrary $\theta \in \mathcal{C}_R \cap X_\psi$. For such a θ, we also have $\theta \in B_R(\psi) = \{\theta | d(\theta, \theta_d(\psi)) < R\}$. Thus $|\theta_j - \theta_i| < 2R \; \forall \; i, j$ (and, in particular, $|x_i| < 2R \; \forall \; i$). These inequalities allow us to find a bound on each $\dot{x}_i$:

$$
\begin{aligned}
\dot{x}_i &= \overline{\dot{\theta_i - \theta_{i+1}}} \\
&= \frac{\alpha}{N}\sum_{j=1}^{N} f(\theta_j - \theta_i) - \frac{\alpha}{N}\sum_{j=1}^{N} f(\theta_j - \theta_{i+1}) \\
&\quad + \frac{\beta}{N-1}h(\theta_i)\sum_{j\neq i}^{N} g(\theta_j) - \frac{\beta}{N-1}h(\theta_{i+1})\sum_{j\neq i+1}^{N} g(\theta_j) \\
&= \frac{\alpha}{N}\sum_{j=1}^{N}[f(\theta_j - \theta_i) - f(\theta_j - \theta_i + x_i)] + \frac{\beta[h(\theta_i) - h(\theta_{i+1})]}{N-1}\sum_{j=1}^{N} g(\theta_j) \\
&\quad + \frac{\beta[h(\theta_{i+1})g(\theta_{i+1}) - h(\theta_i)g(\theta_i)]}{N-1}.
\end{aligned} \tag{4.64}
$$

$$
\dot{x}_i \quad
\begin{cases}
< -\alpha[f'(0)/s]x_i + \beta\frac{N}{N-1}\widehat{h}_1 x_i + \frac{\beta}{N-1}\widehat{G}x_i \triangleq kx_i & \text{if } x_i > 0 \\
> -\alpha[f'(0)/s]x_i + \beta\frac{N}{N-1}\widehat{h}_1 x_i + \frac{\beta}{N-1}\widehat{G}x_i \triangleq kx_i & \text{if } x_i < 0
\end{cases} \tag{4.65}
$$

The inequalities (4.65) use the hypothesis on f', the bound $g(\theta) \leq 1$, and the definitions of $\widehat{h}$ and $\widehat{G}$. Thus, for $k < 0$ (i.e. $\alpha > \alpha_{glob}$), $\dot{\mathcal{R}} = 2\sum_{i=0}^{N-1} x_i \dot{x}_i < 0$ unless $x_i = 0, \forall i$. This argument may be repeated for any ψ and therefore for any arbitrary

$\theta \in \mathcal{C}_R$, so the Proposition follows. □

Since nonlinear stability implies local stability, it must follow from $\alpha > \alpha_{glob}$ that inequality (4.58) is satisfied. This may be seen from the fact that $\alpha > \alpha_{glob}$ implies $\alpha > \alpha_{loc}$ and comparing equation (4.63) with (4.61).

Finally, we note that Proposition 4.4.2 may be sharpened by refining the estimates in (4.65) in any manner that also implies $\text{sign}(\dot{x}_i) = -\text{sign}(x_i)$. For example, a lower value $\widehat{h}_2$ can replace $\widehat{h}_1$ above, where $\widehat{h}_2 = \max_{\theta_i} \widehat{h}_2(\theta_i)$ and $\widehat{h}_2(\theta_i) = \max_\theta \{h'(\theta_i)g(\theta) : |\theta_i - \theta| < 2R\}$ (note that although we have dropped the absolute value in the $\max_\theta$, $\widehat{h}_2 \geq 0$ since h is periodic). The bound $\widehat{h}_2$ arises as follows. If the second term in (4.64) is of opposite sign to x_i, it favors the conclusion $\text{sign}(\dot{x}_i) = -\text{sign}(x_i)$ and hence may be ignored for the purposes of bounding α such that $k < 0$. Thus the natural question is: *assuming* that it is of the same sign as x_i, can we find a smaller upper bound than $\beta \frac{N}{N-1} \widehat{h}_1 x_i$ on the magnitude of this second term? The answer is yes: since $[h(\theta_i) - h(\theta_{i+1})] = [h(x_i + \theta_{i+1}) - h(\theta_{i+1})]$, this difference cannot exceed the upper bound $\beta \frac{N}{N-1} \widehat{h}_2 x_i$, as desired.

4.5 Application to the Hodgkin-Huxley equations

Here we apply the analysis above to study electrotonic and (neurotransmitter driven) synaptic coupling between Hodgkin-Huxley neurons. Existence and stability of clustered states for the Hodgkin-Huxley equations with synaptic coupling has been studied before ([103, 84]), but we contribute results for electrotonic coupling and the strong attraction phase reduction method, as well as a discussion of the stability results in terms of spike time cross-correlograms. First, we introduce the corresponding coupling functions f, g, and h.

4.5.1 Coupling functions

The functions f and g, h, corresponding to electrotonic and synaptic coupling, were computed using both the strong attraction (SA) and phase response curve (PR) methods discussed in Chapter 2. The Hodgkin-Huxley (HH) equations with input current 10 $\mu A/cm^2$ were used ([89]). In their original form these equations were derived from the giant axon of a squid; reduction of mammalian neuron models which include calcium-dependent potassium channels and whose action potential spikes occupy a much smaller fraction of the period than in the HH equations leads to coupling functions somewhat different from those considered here and will be explored in the following chapter.

The effect of electrotonic coupling on the time derivative $\dot{V}_i$ of neuron i's voltage was taken to be $\frac{\alpha}{N} \sum_{j=1}^{N} (V_j - V_i)$ (cf. [101]), and the inhibitory synaptic effect to be $(E_K - V_i) \frac{\beta}{N-1} \sum_{j \neq i} A(V_j, t)$, where E_K is the reversal potential for potassium and $A(V_j, t)$ is an 'alpha function' which takes values in $[0, \tilde{A}]$, $\tilde{A} < 1$, and represents the influence of neuron j on post-synaptic cells. Specifically, $A(V_j, t) = ((t - t_s^j - t_d)/\tau_A) \cdot \exp(-(t - t_s^j - t_d)/\tau_A)$, where t_s^j is the time at which the voltage of neuron j spikes
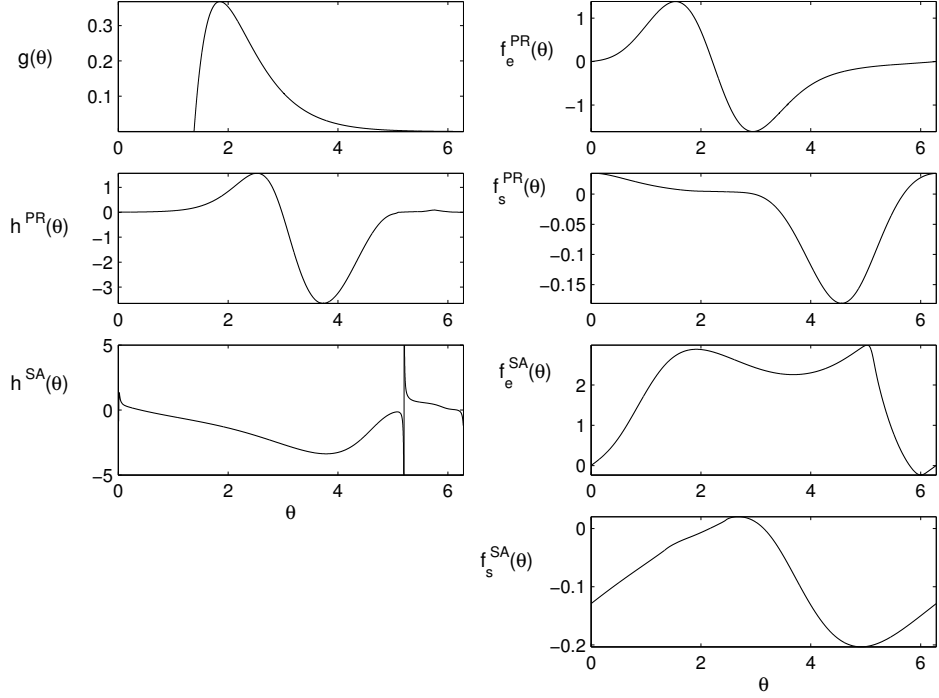
Figure 4.9: Coupling functions derived from the Hodgkin-Huxley equations. The subscript e (s) refers to electrotonic (synaptic) coupling, while the superscript PR (SA) indicates that the function was derived using the phase reduction (strong attraction) method. The h's and g are calculated for the synaptic coupling described in the text: f_s^{PR} is obtained by averaging the product of g and h^{PR}, and f_e^{PR} is obtained by averaging the electrotonic coupling using a phase response method (cf. [103]). f_e^{SA} and f_s^{SA} are obtained by first assuming that the limit cycle is infinitely strongly attracting, followed by averaging (cf. [53, 94]). The 'spikes' in $h^{SA}(\theta)$ are associated with the projection of coupling functions near turning points in the original phase variables; the tips extend to approximately ± 30 (there is also an $\mathcal{O}(1)$ spike near $\theta = 0$, not visible here).

(see below), t_d is the synaptic delay, and τ_A is the synaptic time constant (e.g. [103]). We take $I_b = 10 \ \mu A/cm^2$ in the HH equation so that their period is $T = \frac{2\pi}{\omega}$ is 14.64 $msec$, and take $\tau_A = 1.1 \ msec$ and $t_d = 6.6 \ msec$. These neuron and coupling models and parameters lead to the reductions of the coupling functions to T^N displayed in Figure 4.9.

4.5.2 Phase-difference coupling

In this section we assume that synaptic coupling $\frac{\beta}{N-1} h(\theta_i) \sum_{j \neq i} g(\theta_j)$ is sufficiently weak that it can be averaged to yield $\frac{\beta}{N-1} \sum_{j \neq i} f_s(\theta_j - \theta_i)$, and we take $\beta = \kappa\alpha$, so that (4.50) becomes $\dot{\theta}_i = \omega + \frac{\alpha}{N} \sum_j f(\theta_{ji})$, where the phase difference function $f(\cdot) = f_e(\cdot) + \kappa f_s(\cdot)$. We use the methods of Section 4.3 to determine stability of

	PR $\kappa = 0$	PR $\kappa = 10$	PR $\kappa = 50$	PR $\kappa = 100$	SA $\kappa = 0$
$(\mathbf{S_k})^\mathbf{m}\dotplus\mathbf{Z_m}$					
S, $m =$	1,2,4	1,2,4	1,4	1	1,3,4
U, $m =$	3,6,8,12	3,6,8,12	2,3,6,8,12	2,3,4,6,8,12	2,6,8,12
$\mathbf{S_{N-p}} \times \mathbf{S_p}$					
S, $p =$	11,12	12	1,2	1-4	1-6
U, $p =$	1-12	1-12	1-12	3-12	7-12

Table 4.1: The linear stability of various 'clustered' periodic orbits of diagonal flow for different phase difference couplings $f_e(\theta_{ji}) + \kappa f_s(\theta_{ji})$, computed using the PR or SA methods. $(S_k)^m \dotplus Z_m$ stability is given for allowable m ($N = 24$), and $S_{N-p} \times S_p$ stability for $N = 24$ and $p = 1, ..., 12$; a value of p being listed twice indicates the correspondence of multiple δ's. S and U indicate asymptotic stability and instability, respectively, for the m or p values given in the subsequent columns.

periodic orbits. The results are shown in Table 4.1 for various values of κ and coupling functions derived with the PR and SA methods. The SA coupling functions give rise to a larger set of distinct stable periodic orbits, the consequences of which will be discussed below.

To check the validity of the SA and PR phase reductions, we compared a few cases of our stability predictions with numerical simulations of the full HH equations (cf. [103]); for $\kappa = 0$, stability was consistent with the full equations for the PR reductions but *not* for SA. However, to illustrate properties of (4.8) we will continue to refer to the SA functions when their general form gives additional (e.g. contrasting) results.

4.5.3 Phase-dependent coupling

We now return to consider equation (4.50) with unaveraged synaptic (product) coupling. Since $f(0) = 0$ for electrotonic coupling, we may use equation (4.62) to explicitly calculate a lower bound on α for this orbit to be stable. The resulting $\tilde{\alpha}_{loc} < 0$ (Figure 4.11), so that the in-phase state is stable for synaptic coupling and any positive α (since $f'(0) > 0$). In addition, the domain of attraction may be estimated: for example, taking $s = 2.14$ in Proposition 4.4.2 gives $R = 1/4$ and $\hat{h}_2 = 0.64$ for PR coupling functions with $\kappa = 0$. Thus, the domain of attraction of the in-phase orbit includes $\mathcal{C}_R$ if $\alpha > \alpha_{glob} = 7.11\beta$ (equation (4.63) for $N \to \infty$). Figure 4.10(d) demonstrates the collapse of the cross correlogram (b) upon addition of the synchronizing (cf. Section 4.5.3) product coupling.

For the SA coupling functions, the following observation is useful: since the denominator in the integrand of (4.58) is always positive, if $h'(\theta) < 0$ for θ in the (essential) support of the positive function g then the integrand itself will always be *negative*, giving stability for any $\alpha > 0$. The plots in Figure 4.9 show that for delays t_d (which correspond to translations of g) taken in a wide range around 6.6 *msec*, stability holds for arbitrary g of reasonably compact essential support and *any* $\alpha, \beta > 0$ such

that the in-phase periodic orbit exists. We note here that related stability conditions are derived in [171, 50, 69], in which the time course of g is also important.

4.5.4   Cross correlograms

This section discusses the qualitatively distinct cross correlograms patterns that can arise under different coupling functions. Cross correlograms are derived from solutions of (4.50) as follows. A spike is deemed to occur when a rotator θ_i crosses through a threshold value θ_s: the solution of $\{V(\theta_s) = V_s, V'(\theta_s) > 0\}$, where $V_s = -30mV$ is a depolarized voltage characteristic of a neuron firing an action potential and the function $V(\theta)$ is defined by $V(t) = V(\theta/\omega)$ over the period of one neuron action potential. The set of all pairwise differences between times at which distinct spike events occur is computed according to this definition, and the cross correlogram is the histogram of this set.

To produce illustrative correlograms, we consider equation (4.8) in the presence of noise represented as additive Brownian forcing on the torus, so that:

$$d\theta_i = \left[\omega + \frac{\alpha}{N}\sum_{j\neq i}^{n} f\left(\theta_j - \theta_i\right)\right]dt + \sigma dW_t^i. \tag{4.66}$$

The inclusion of random noise represents additional input currents, a common stratagem in accounting for the influence of neural subgroups neglected in the model. Stochastic averaging leads to the approximation above, as discussed in Chapter 3, Section 3.5. Simulations of equation (4.66) to be discussed below were performed using a second order stochastic Runge-Kutta method ([90]).

For phase-difference coupling, the stability results of Table 4.1 will persist for α sufficiently larger than σ, but solutions approach a constant-drift Brownian motion on T^N as the coupling-to-noise ratio α/σ decreases. Figure 4.10(a) shows a representative flat cross-correlogram in this small coupling-to-noise regime: there is no preferred firing time difference between neurons. However, for the PR coupling functions in which there are relatively few stable periodic orbits significantly different from the in-phase mode (in particular, $\delta < 1$ for the stable $S_p \times S_{N-p}$ modes for all κ values in Table 4.1), as the coupling strength α is increased, a broad central peak (with intermediate 'shoulders') emerges (Figure 4.10(b)).

If a variety of diagonal flow solutions are simultaneously stable (as for the SA coupling functions) the following mechanism can produce broad 'peak-shoulder' cross correlogram patterns for much lower values of noise. For every revolution (of diagonal flow), $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$ states with $\delta \neq 0, \pi$ produce cross correlograms with $m(k_1^2 + k_2^2 - k_1 - k_2)$ counts at $t = 0$, $m(k_1^2 + k_2^2)$ counts at times proportional to $\pm 2\pi j/m$, $j = 1, ..., m-1$, and mk_1k_2 counts at times proportional to $\pm(2\pi j/m) \pm \delta$, $j = 0, ..., m-1$. If $N \geq 5$, this leads to a dominant central peak in the cross correlogram for (two-cluster) states with $m = 1$. Moreover, all peaks *except* for the central peak at 0 will be differently spaced for each distinct $(S_{k_1} \times S_{k_2})^m \dot{+} Z_m$ orbit. Thus, if the individual cross correlograms from many of these states are combined (e.g. due to stochastic switching due to random noise in (4.66)), the common central maxima can
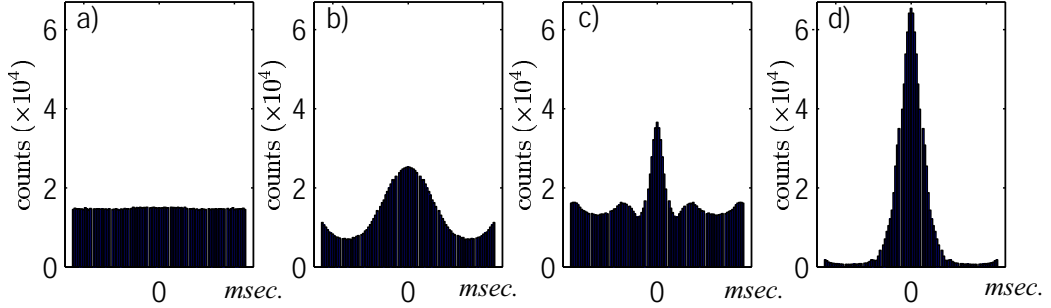
Figure 4.10: Cross correlograms for simulations of the phase-difference model (a-c). To facilitate comparison, parameters $\bar{\alpha}_{PR}$ and $\bar{\alpha}_{SA}$ were first chosen so that both $\bar{\alpha}_{PR} \max_\theta\{|f_e^{PR}(\theta)|\} = 2.25$ and $\bar{\alpha}_{SA} \max_\theta\{|f_e^{SA}(\theta)|\} = 2.25$. (a) PR functions with low coupling-to-noise $\alpha = \frac{1}{5}\bar{\alpha}_{PR}$, $\sigma = 0.8$, and $\kappa = 0$; (b) same except higher coupling-to-noise via $\alpha = \bar{\alpha}_{PR}$. (c) SA coupling functions with $\alpha = \bar{\alpha}_{SA}$ and the lower noise value $\sigma = 0.2$ (again with $\kappa = 0$). (d) as in (b), but with the addition of phase-dependent (synaptic) coupling of strength $\beta = 0.23$. The range of all histograms is $[-.7T, .7T]$, where $T = 14.64$ $msec$ is the period of the HH equations; all histograms are averaged over five simulated recordings with uniformly distributed initial conditions with $N = 24$.

conspire to produce a central peak in the cross correlogram while the combination of many secondary maxima could give rise to the relatively flat shoulder. This is demonstrated in Figure 4.10(c).

4.5.5 Frequency effects

In this section we study how the neural coupling functions affect frequency of the in-phase state.

Phase-difference coupling

The period of the in-phase state of (4.66) with phase-difference coupling only is $\frac{2\pi}{\omega + \alpha f(0)}$, and Figure 4.9 shows that $f(0) > 0$ for any $\kappa > 0$, so that the period will always *decrease* as α increases.

Phase-dependent coupling

For small β, where averaging is valid, the period of the S_N symmetric orbit must decrease with β. However, Figure 4.11 also shows that for β sufficiently large, the period *increases* with β (the term $G(\theta)$ in (4.51) slows the flow). Since the S_N symmetric orbit is attracting, its increasing period indicates a mechanism for lower firing rates with sufficiently increased coupling. This shows the importance in this case of considering the explicit product coupling of (4.50) for the HH neuron; however,
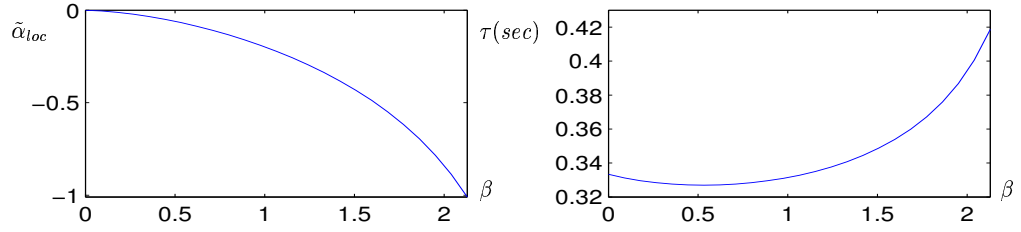
Figure 4.11: The β-dependence of (left) bounds for stability and (right) the period of the S_N symmetric orbit with PR coupling functions, from equations (4.62) and (4.57) in the large-N limit.

for other parameter values and neuron models phase-difference coupling does correctly capture trends in firing rates (cf. [24] and Chapter 5).

Appendix: eigenvalue calculations

Eigenvalues for rotating block solutions

Lemma 4.5.1. *Define the set $\mathcal{M} = \{l | l = qm \text{ for some } q \in \mathbb{Z}\}$ and let $\gamma = \exp(2\pi i/m)$. Then for $\bar{l} \in \mathbb{Z} \setminus \mathcal{M}$, $\sum_{r=1}^{m-1} \gamma^{\bar{l}r} = -1$.*

Proof. Let

$$A = \sum_{r=0}^{m-1} \gamma^{\bar{l}r} \quad \text{and} \quad B = \sum_{r=1}^{m-1} \gamma^{\bar{l}r} \ ,$$

then

$$B = A - 1 \ . \tag{4.67}$$

Also, using $\gamma^{m\bar{l}} = 1$, note that

$$\gamma^{\bar{l}}B = \gamma^{2\bar{l}} + ... + \gamma^{m\bar{l}} = A - \gamma^{\bar{l}} \ , \tag{4.68}$$

Upon subtracting Eqns. (4.67) and (4.68), we have

$$B\left(1 - \gamma^{\bar{l}}\right) = -1 + \gamma^{\bar{l}} \Rightarrow B = -1 \ , \tag{4.69}$$

provided $\gamma^{\bar{l}} \neq 1$, which is guaranteed if $\bar{l} \notin \mathcal{M}$. $\qquad\square$

Proof of Proposition 4.3.5. Linearized about $\bar{\phi}$, Eqns. (4.8) become

$$\dot{\xi}_i = -\frac{\alpha}{N}\sum_{j=1}^{N} f'(\bar{\phi}_{ji})\xi_i + \frac{\alpha}{N}\sum_{j=1}^{N} f'(\bar{\phi}_{ji})\xi_j \triangleq \left[J(\bar{\phi})\xi\right]_i \ . \tag{4.70}$$

The Jacobian has m^2 blocks of size $(k \times k)$ [6]; the 2π periodicity of f' ensures that there are only $m+1$ distinct entries. The Jacobian takes the form

$$J(\bar{\phi}) = \frac{\alpha}{N}\begin{bmatrix} D & M_1 & \cdots & M_{m-1} \\ M_{m-1} & D & \cdots & M_{m-2} \\ \cdots & & & \\ M_1 & M_2 & \cdots & D \end{bmatrix}, \tag{4.71}$$

where

$$D \triangleq \begin{bmatrix} c_{-1} & c_0 & \cdots & c_0 \\ c_0 & c_{-1} & \cdots & c_0 \\ \cdots & & & \\ c_0 & c_0 & \cdots & c_{-1} \end{bmatrix}, \qquad M_j \triangleq \begin{bmatrix} c_j & c_j & \cdots & c_j \\ c_j & c_j & \cdots & c_j \\ \cdots & & & \\ c_j & c_j & \cdots & c_j \end{bmatrix},$$

are $(k \times k)$ blocks, and

$$c_j = f'\left(\frac{2\pi j}{m}\right), \quad j = 0, \cdots, m-1, \tag{4.72}$$

$$c_{-1} = -\left[(k-1)c_0 + k\sum_{r=1}^{m-1} c_r\right]. \tag{4.73}$$

Rotation eigenvalues. The shifted row structure of $J(\bar{\phi})$ classifies it as a (block) circulant or Toeplitz matrix. Such matrices have the eigenvectors (e.g. [6]):

$$v_j = (\gamma^{0j}, \gamma^{0j}, ..., \gamma^{0j}, ..., ..., \gamma^{(m-1)j}, \gamma^{(m-1)j}, ..., \gamma^{(m-1)j})^{\mathrm{T}} \quad j = 0, ..., m-1 \;, \tag{4.74}$$

where $\gamma = \exp(2\pi i/m)$ and each element in v_j is repeated k times. For the special form of $J(\bar{\phi})$ in Eqn. (4.71), substitution of (4.74) into the eigenvalue problem gives the eigenvalues corresponding to the (real or imaginary parts of the) v_j:

$$\lambda_j \frac{N}{\alpha k} = \sum_{r=1}^{m-1} c_r(\gamma^{jr} - 1) \;. \tag{4.75}$$

Clearly $\lambda_0 = 0$ with $v_0 = (1, 1, \ldots, 1)$, corresponding to the T^1 invariance. The eigenvalues for $j > 0$ are the 'rotation eigenvalues,' and may be found in terms of the series (4.9) using

$$c_r = \sum_{l=1}^{\infty} l\left(b_l^o \cos\left(\frac{2\pi rl}{m}\right) - b_l^e \sin\left(\frac{2\pi rl}{m}\right)\right) \;, \quad r = 1, \ldots, m-1 \;. \tag{4.76}$$

Eqns. (4.75), (4.76) then give

$$\lambda_j \frac{N}{\alpha k} = \sum_{r=1}^{m-1}\sum_{l=1}^{\infty} l\left(b_l^o \cos\left(\frac{2\pi rl}{m}\right) - b_l^e \sin\left(\frac{2\pi rl}{m}\right)\right)(\gamma^{jr} - 1) \;. \tag{4.77}$$

Upon switching the order of the summation, the term corresponding to each l in the sum of Eqn. (4.77) may be written as

$$\frac{l}{2}\sum_{r=1}^{m-1}\left(b_l^o\left(\gamma^{rl} + \gamma^{-rl}\right) + ib_l^e\left(\gamma^{rl} - \gamma^{-rl}\right)\right)(\gamma^{jr} - 1) =$$

$$\frac{b_l^o l}{2}\left(\sum_{r=1}^{m-1}\left(\gamma^{r(j+l)} + \gamma^{r(j-l)}\right) - \sum_{r=1}^{m-1}\left(\gamma^{rl} + \gamma^{-rl}\right)\right)$$

$$+\frac{ib_l^e l}{2}\left(\sum_{r=1}^{m-1}\left(\gamma^{r(j+l)} - \gamma^{r(j-l)}\right) - \sum_{r=1}^{m-1}\left(\gamma^{rl} - \gamma^{-rl}\right)\right) \equiv$$

$$\frac{b_l^o l}{2}(I - II) + \frac{ib_l^e l}{2}(III - IV). \tag{4.78}$$

We will consider separately the real and imaginary terms in Eqn. (4.78). By Lemma 4.5.1, for l satisfying (i) $l \in \mathbb{Z}^+ \setminus \mathcal{M}$, $II = -2$. Similarly, for l meeting (ii) $j + l \in \mathbb{Z} \setminus \mathcal{M}$ and (iii) $j - l \in \mathbb{Z} \setminus \mathcal{M}$, $I = -2$. Thus, for values of l satisfying all three conditions (i-iii), $I - II = 0$ and the corresponding term in the Fourier expansion for f makes no contribution to the real part of λ_j.

Now we determine the values of l for which the conditions do not hold and hence which terms contribute to the real part of the sum in Eqn. (4.77). Obviously, (i) does not hold where $l \in \mathcal{M}$. Conditions (ii) and (iii) are violated for values of l such that, resp.,

$$l = mh_1 - j \text{ or } l = mh_2 + j \text{ for some } h_1, \ h_2 \ \in \mathbb{Z} . \tag{4.79}$$

Recalling that $l \in \mathbb{Z}^+$ and $j \in \{1, ..., m-1\}$, the conditions (ii) and (iii) are only met when $h_1 \in \{1, 2, ...\}$ and $h_2 \in \{0, 1, 2, ...\}$, resp. Thus, from the definition of the sets $\mathcal{M}_1^j$, $\mathcal{M}_2^j$, and $\mathcal{M}$, we may conclude that the only nonzero real terms in the expression for λ_j come from values of l in these sets. Furthermore, the fact that $j \in \{1, ..., m-1\}$ also implies the that (i) may not be violated simultaneously with (ii) and/or (iii) (where simultaneously implies for a single pair (l, j)). In other words, $\{\mathcal{M}_1^j \bigcup \mathcal{M}_2^j\} \bigcap \mathcal{M} = \emptyset$.

The nonzero real terms may therefore be calculated as follows. For $l \in \mathcal{M}$, $\gamma^{rl} = 1$ so $II = 2(m-1)$. Since $l \in \mathcal{M}$ implies $l \notin \mathcal{M}_{1,2}^j$, in this case $I = -2$ so that $I - II = -2m$. For $l \in \{\mathcal{M}_1^j \bigcup \mathcal{M}_2^j\} \setminus \{\mathcal{M}_1^j \bigcap \mathcal{M}_2^j\}$, by the lemma and disjointness of $\mathcal{M}$ and $\mathcal{M}_{1,2}^k$ we have $II = -2$, $I = m - 2$, so that $I - II = m$. Similarly, $l \in \{\mathcal{M}_1^j \bigcap \mathcal{M}_2^j\}$ gives $II = -2$ and $I = 2(m-1)$, or $I - II = 2m$. Recalling that $N = mk$, this proves the real part of Eqn. (4.28) and in particular indicates that the conditions $l \in \{\mathcal{M}_1^j \bigcup \mathcal{M}_2^j\} \setminus \{\mathcal{M}_1^j \bigcap \mathcal{M}_2^j\}$ and $l \in \{\mathcal{M}_1^j \bigcap \mathcal{M}_2^j\}$ need not be distinguished in the sum.

For the imaginary terms, we first note that Lemma 4.5.1 implies $IV = 0$ for any l. Similarly, when $l \notin \{\mathcal{M}_1^j \bigcup \mathcal{M}_2^j\} \setminus \{\mathcal{M}_1^j \bigcap \mathcal{M}_2^j\}$, $\sum_{r=1}^{m-1} \gamma^{r(j+l)} = \sum_{r=1}^{m-1} \gamma^{r(j-l)}$ so that $III = 0$. For $l \in \{\mathcal{M}_2^j \setminus \mathcal{M}_1^j\}$, $\sum_{r=1}^{m-1} \left(\gamma^{r(j+l)} - \gamma^{r(j-l)}\right) = -1 - (m-1) = -m$, and for $l \in \{\mathcal{M}_1^j \setminus \mathcal{M}_2^j\}$, $\sum_{r=1}^{m-1} \left(\gamma^{r(j+l)} - \gamma^{r(j-l)}\right) = (m-1) + 1 = m$. Again using $N = mk$, this gives the imaginary term of Eqn. (4.28).

Permutation eigenvalues. Inserting $\lambda_p = \alpha(c_{-1} - c_0)$ into the eigenvalue equation $(J(\bar{\phi}) - \lambda I)v$ gives a linear system with m blocks of identical elements c_0 on the diagonal. Thus any vector of the form

$$(0, \ldots, 0, a_1, \ldots, a_k, 0, \ldots, 0)^{\mathrm{T}},$$

where a_1 is in the $(nk + 1)$st place, $n \in \mathbb{Z}$ and $\sum_{p=1}^{k} a_p = 0$, is an eigenvector with eigenvalue λ_p. The latter equation has $k - 1$ solutions for each of the m possible positions of a_1, so that λ_p has multiplicity $m(k-1)$. To compute λ_p in terms of the Fourier coefficients of f, note that

$$\lambda_p \frac{N}{\alpha} = c_{-1} - c_0 = -k \sum_{r=0}^{m-1} c_r . \tag{4.80}$$

Switching the order of the summation over Fourier modes and using Eqn. (4.76) we find

$$\lambda_p \frac{N}{\alpha} = \sum_{l=1}^{\infty} \left(\frac{-kb_l^o l}{2} \sum_{r=0}^{m-1} \left(\gamma^{lr} + \gamma^{-lr} \right) - \frac{ikb_l^e l}{2} \sum_{r=0}^{m-1} \left(\gamma^{lr} - \gamma^{-lr} \right) \right) . \tag{4.81}$$

Lemma 4.5.1, the fact that $\gamma^0 = 1$, and the above remark that $IV = 0$ now imply the proposition. $\qquad\square$

Eigenvalues for two-block solutions

Ordering the oscillators so that $\phi_{ji} = 0$ for the groups having $\{i, j\} \in \{1, ..., p\}$ and $\{i, j\} \in \{p + 1, ..., N\}$, the linearization (4.70) becomes

$$\dot{\xi} = A_2 \xi , \tag{4.82}$$

where for $1 \le p \le N - 1$

$$A_2 \equiv \frac{\alpha}{N} \begin{bmatrix} N_1 & a & \cdots & a & -b & -b & \cdots & -b \\ a & N_1 & \cdots & a & -b & -b & \cdots & -b \\ & & \cdots & & & & & \\ a & a & \cdots & N_1 & -b & -b & \cdots & -b \\ -c & -c & \cdots & -c & N_2 & a & \cdots & a \\ & & \cdots & & & & & \\ -c & -c & \cdots & -c & a & a & \cdots & N_2 \end{bmatrix} . \tag{4.83}$$

Here, $N_1 = (N - p)b - (p - 1)a$, $N_2 = pc - (N - p - 1)a$, and the diagonal subblocks are of dimension $p \times p$ and $(N - p) \times (N - p)$.

Inserting $\lambda_1 = \frac{\alpha}{N}(N_1 - a)$ into the eigenvalue equation $(A_2 - \lambda_1 I)v = 0$ gives

$$(A_2 - \lambda_1 I)v = \frac{\alpha}{N} \begin{bmatrix} a & a & \cdots & a & -b & -b & \cdots & -b \\ a & a & \cdots & a & -b & -b & \cdots & -b \\ & \cdots & & & & & & \\ a & a & \cdots & a & -b & -b & \cdots & -b \\ -c & -c & \cdots & -b & d & a & \cdots & a \\ & \cdots & & & & & & \\ -c & -c & \cdots & -b & a & a & \cdots & d \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0 , \tag{4.84}$$

where $d = N_2 - N_1 + a$. Eqn. (4.84) is solved by $p - 1$ linearly independent eigenvectors as shown and such that $v_1 + v_2 + ... + v_p = 0$. The proof of the λ_2 results is similar, and those for λ_3 follow from the fact that each row of (4.84) sums to 0. Adding λ_1 through λ_3 with multiplicities and comparing the result with $\mathrm{Tr}(A_2) = \alpha \left(pN_1 + (N - p)N_2 \right)$ yields the formula for λ_4. The case $p = 0$ is treated similarly.

Chapter 5

Firing rates and synchrony in the locus coeruleus

5.1 Chapter outline

In this chapter we apply the results of Chapters 2-4 to spiking neurons in *locus coeruleus* (LC), a brain nucleus involved in modulating cognitive performance, and compare with recent experimental data. Extracellular recordings from LC of monkeys performing target detection and selective attention tasks show varying responses dependent on stimuli and performance accuracy. From membrane voltage and ion channel equations, we derive a phase oscillator model for LC neurons using the asymptotic phase reduction method discussed in Chapter 2. Average spiking probabilities of a pool of cells over many trials are then computed via a probability density formulation, and applying the analysis of Chapter 3 then shows that: 1) Post-stimulus LC response is elevated in populations with lower spike rates; 2) Responses decay exponentially due to noise and variable pre-stimulus spike rates; and 3) Shorter stimuli preferentially cause depressed post-activation spiking. We also derive coupling functions for the (presumed weak) synaptic and electrotonic connections between LC neurons, and use these with the methods of Chapter 4 to model and explain empirical findings on synchrony in the LC. These results allow us to propose mechanisms for the different LC responses observed across behavioral and task conditions, and to make explicit the role of baseline firing rates and the duration of task-related inputs in determining LC response. The paper [24] is an elaborated version of this Chapter.

5.2 Introduction and background

The *locus coeruleus* (LC) is a brainstem nucleus containing approximately 15,000 neurons in monkeys (35,000 in humans), each of which can make 100,000 or more synapses with its widespread target regions, including the cerebral cortex [123, 62]. LC neurons release norepinephrine, which is known to modulate brain processes including the sleep/wake cycle and arousal [62, 8]. Recent data indicate that the LC regulates attention and behavioral flexibility [10, 167, 9]. Specifically, increased tran-

sient LC activity may increase the responsivity of decision networks following salient stimuli, hence improving accuracy and response time. (These effects are investigated computationally in Part II of this dissertation.) Conversely, lower baseline LC activity would reduce responsiveness to distractors [167, 154].

Neurons in the alert monkey LC exhibit two distinct modes: *phasic* and *tonic* [78, 10, 167]. In the latter, associated with labile behavior and poor performance on tasks requiring focused attention, cells fire at relatively high rates with little synchrony; in the former, associated with good performance, firing rates are lower but display greater synchrony. The phasic mode also produces greater response to stimuli, as detailed below. We emphasize that the phasic and tonic modes are defined based on good vs. poor task performance alone, but that this performance correlates strongly with baseline LC firing rates, as detailed below (also see Fig. 1A of [167]).

Two previous modeling studies have proposed mechanisms for the different firing properties of the phasic and tonic modes. The computational model of [167] used a pool of coupled integrate-and-fire neurons and found, via numerical simulation, that increased electronic coupling promotes synchrony and enhanced responses to task stimuli. More recently, experimental and computational studies of paired *in vitro* LC cells have shown that decreased baseline activity can enhance the synchronizing effects of fixed-strength electrotonic coupling [1], cf. [32]; however, the different patterns of LC response to stimuli in the phasic vs. tonic modes were not addressed in these studies. Here, we show how decreased baseline spiking can, via different mechanisms, cause not only the partial synchronization but also the amplified response to exogenous inputs observed in the phasic mode. Hence we propose decreased baseline spiking rates in the phasic mode, resulting from altered exogenous input to the LC, as a new mechanism contributing to the phasic/tonic transition (as elaborated in the Discussion, this does not exclude other effects such as coupling). We suggest that these lower rates may result from decreased excitatory or enhanced inhibitory input from brain areas afferent to the LC (including the anterior cingulate cortex (ACC), a region previously implicated in cognitive control) or from neuromodulators such as corticotropin releasing factor (CRF). Neural evidence for these possibilities is elaborated upon in the Discussion.

Recent data indicates that LC responses differ not only between LC modes, as just discussed, but also among different psychological tasks. In Section 5.3, we present data demonstrating this difference for the target identification vs. Eriksen flanker tasks. This data also indicates different LC responses in trials in which correct vs. incorrect behavioral responses to task stimuli were obtained. In this chapter, we show how differences in LC responses among task types and behavioral conditions can be accounted for by assuming different time-courses of inputs to the LC in the two tasks, as well as greater variation of input arrival times in incorrect vs. correct conditions, in accord with behavioral data. Indeed, these different LC inputs are a *prediction* of our model.

Below we develop a mathematical model for a pool of LC neurons, reduce it to differential equations for individual neuron phases, thereby retaining spike timing information, and analyze spiking probabilities in response to stimuli. This elucidates the dependence of spike histograms on model parameters and reveals how timescales

in the neural substrate interact with those in the stimuli. Section 5.3 reviews relevant experimental data and provides an overview of its relationship with the modelling results that follow. We describe the neural model and probabilistic analysis in Sections 5.4 and 5.5, fit parameters and compare model results with data in Section 5.6, and discuss the work in Section 5.7.

5.3   Experimental data and modelling overview

5.3.1   Experimental results and methods

This section summarizes experiments carried out in the laboratory of G. Aston-Jones (Dept. of Psychiatry, Univ. of Pennsylvania), who kindly made his data available to us.

Activities of individual neurons were obtained from behaving monkeys using extracellular recording techniques, as described previously [10]. Animals were trained to continuously depress a pedal and visually fixate a centrally located spot on a video monitor. In the target detection task, after successful fixation, target (20% of trials) or non-target cues (80% of trials) were displayed singly in random order across trials, with random inter-trial intervals (1.65 sec on average). Release of the pedal within 650 msec after target cue onset was rewarded by juice. Four response categories are possible: correct detection (hit), correct rejection, incorrect detection (false alarm), and incorrect omission (miss).

The second task, the Eriksen flanker paradigm [47], requires greater attentional focus. The display comprises five icons, with two 'flankers' on each side of the central cue, each selected at random with 50% probability from two possible cues ('left' or 'right') . The subject was trained to respond by releasing the left or right of two pedals according only to the central icon. The distracting flankers were either all identical to the central cue (congruent stimulus) or identical to the opposite, non-displayed cue (incongruent stimulus).

Extracellular recordings from LC neurons were obtained from microwire electrodes positioned within the brain via a stereotaxically implanted guide cannula. LC neurons were identified during recording sessions by electrophysiological criteria, and continuous monitoring of eye position and pupil diameter was performed, as previously described [10]. Baseline activity was calculated as an average spike rate during 500 msec epochs immediately preceding stimuli. Peri-stimulus time histograms (PSTHs) were produced and population PSTHs generated by aligning visual stimulus onsets and averaging across multiple sessions, or selected portions thereof. Histograms were smoothed via averages of spike counts in neighboring bins (using the program SigmaPlot) to facilitate superposition of cumulative PSTHs in a single figure while preserving response pattern and timing.

Fig. 5.1 shows examples of the resulting PSTH data. These histograms reveal LC responses to stimuli for both tasks. As previously reported [167], in the target detection task, response relative to baseline is greater during good (phasic LC mode) compared to poor (tonic LC mode) performance, and a period of depressed spiking
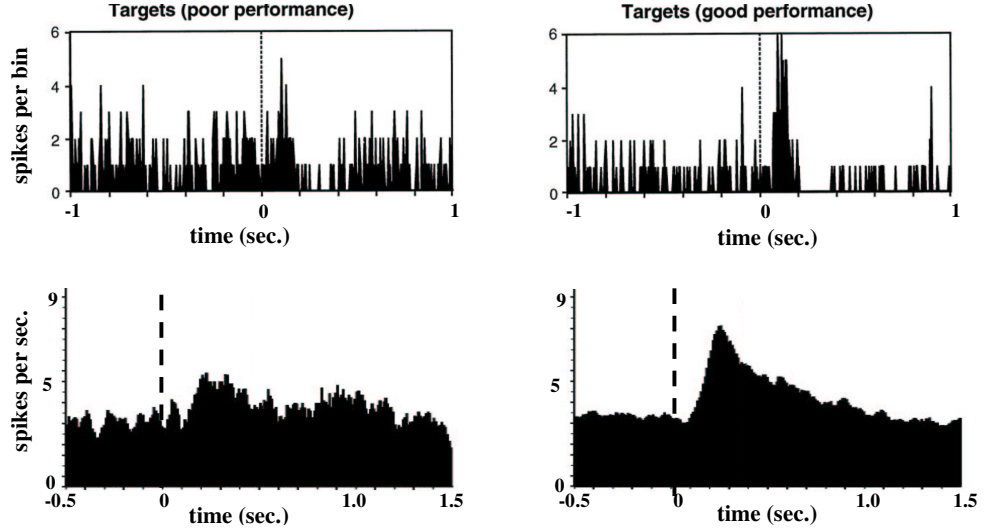
Figure 5.1: Peri-stimulus time histograms of LC activity for poor (left) and good (right) performance periods during the target identification task (top) and for incorrect (left) and correct (right) responses in the Eriksen task (bottom). Former are from single neurons, averaged over ~ 100 trials; latter are cumulative PSTHs from multiple neurons averaged over ~ 600 and 6000 trials respectively. Histograms are aligned at visual stimulus onset, marked by dashed line.

follows the response before activity returns to baseline in both cases. The phasic mode also displays greater synchrony; cf. Fig. 5.8 below. Reduced spiking following the LC response is *not* observed for the Eriksen task; instead, near-monotonic decay occurs following phasic activation, and the magnitude is much reduced for incorrect responses. Furthermore, in this latter task pre-stimulus spike rates are similar for both correct and incorrect responses [33]; the animal seems not to undergo phasic/tonic transitions, although baseline frequencies do show slow variations.

5.3.2   Assumptions and modelling overview

Fig. 5.1 shows three ways in which LC response differs among task and behavioral conditions: between the phasic and tonic modes in the target detection task, between the target detection and the Eriksen flanker tasks, and between correct and incorrect trials in the Eriksen task. In this chapter we develop a model of LC response to stimuli and use it to propose mechanisms for these differences. Our model is based on the following assumptions: (A1) Different levels of baseline current input to LC neurons determine the different distributions of baseline firing rates reported in [167] (p. 550, col. 1) for the phasic (slower, more tightly distributed rates: 2 ± 1.1 Hz, (mean $\pm$ st. dev.)) vs. tonic (faster, more broadly distributed: 3 ± 1.6 Hz) modes. (Note that these values are reported in [167] as mean $\pm$ SEM.) (A2) In contrast to [167] and to clearly separate the effects of baseline firing rate, coupling strengths are chosen to be identical in the phasic vs. tonic modes. (A3) The strength and duration

91

of stimulus-related input to the LC may differ between the target detection and Eriksen tasks, as described in Section 5.4.3. (A4) Reflecting the greater variability in reaction times on incorrect vs. correct trials, onsets of stimulus-evoked inputs to the LC may also be more variable for incorrect trials. This latter assumption is justified under the hypothesis that pulsed inputs to the LC are driven by the (stochastic) decision process [167]; more-variable timing of responses (reaction times) then implies increased variability of "triggering times" for LC inputs.

Analysis of our model yields three main mathematical results, presented in bold-face in the text of Section 5.5. These are (1) Maximum LC response is elevated in populations with slower baseline firing rates, (2) response decays exponentially or faster with t due to noise and heterogeneous frequencies, and (3) in systems with narrow frequency distributions, short inputs necessarily lead to intervals of depressed firing following enhanced spiking and stimulus offset.

The first of these findings explains the influence of baseline spike rate on response of LC neurons to exogenous stimuli, and suggests that any factor leading to decreased baseline rate contributes to stronger responses in the phasic vs. tonic mode; see Fig. 5.5 (top right panel vs. bottom left) and Fig. 5.9 (top). The second finding tells us that, in order to produce the protracted LC responses seen in the Eriksen data, Eriksen task stimuli must elicit protracted inputs to the LC (since the impact of brief inputs decays quickly). The third finding implies that, in contrast to the Eriksen task, inputs to the LC elicited by target detection task stimuli must be punctate, because the interval of depressed firing observed in the data (Fig. 5.1, top) can occur only following the offset of brief (pulsed) inputs to the LC. Therefore the second and third findings address influence of stimulus duration on response of LC neurons. We also show via additional simulations that varying onset times of stimulus-evoked LC inputs in incorrect vs. correct Eriksen trials (in proportion to reaction time variability under these conditions) reproduces the trend in Fig. 5.1 (bottom).

5.4 A mathematical model for LC neurons

5.4.1 A conductance-based model and phase reduction

LC neurons possess calcium- and voltage-dependent potassium currents (e.g., 'A-currents'), which are largely responsible for their slow ($\lesssim 8$ Hz) firing rate [179] and their resulting classification as 'Type I' cells [50]. We base our model of individual LC neurons on the original model by [41] (cf. [149]) for a multi-ion-channel 'Type I' neuron including the A-current. This is a generic choice, intended to capture the essence but not necessarily the biophysical detail of LC dynamics. For ease of computation, we exploit a further simplification by Rose and Hindmarsh [146], who used differing timescales and approximate relationships among state variables to reduce the Connor

et al. system to two variables:

$$\dot{V}_i = [I_i^b - g_{Na}m_\infty(V_i)^3(-3(q_i - Bb_\infty(V_i)) + 0.85)(V_i - V_{Na}),$$
$$-g_K q_i(V_i - V_K) - g_L(V_i - V_L) + I_i^{ext}]/C \qquad (5.1)$$
$$\dot{q}_i = (q_\infty(V_i) - q_i)/\tau_q(V_i) .$$

Rose and Hindmarsh show that this reduction compares favorably with the original system. Here V_i is the voltage of neuron i ($i = 1, \ldots, N$ for an N-cell model), q_i is a collective gating variable, C is cell membrane conductance, g_{Na}, g_K, and g_L are maximum conductances for sodium, potassium, and leak currents, and I_i^b is the baseline inward current, which effectively sets spike frequency. I_i^{ext} denotes extracellular currents described below, and the other terms are channel gating variables, $m_\infty(V)$ and $b_\infty(V)$ denoting equilibrium levels for fast sodium and potassium channels. Functional forms and parameter values for this Rose-Hindmarsh model are given in the appendix to Part I of this dissertation.

LC neurons are coupled by: 1) voltage differences between cells in electrical contact at electrotonic or gap junctions; 2) neurotransmitter release across synaptic clefts following presynaptic spikes. These effects, along with currents $I(t)$ representing inputs due to external stimuli, enter I_i^{ext}:

$$I_i^{ext} = \frac{\beta_e}{N} \sum_{j=1}^{N}(V_j - V_i) + \frac{\beta_s}{N} \sum_{j=1}^{N}\sum_{k} A(t - t_{jk})(V_K - V_i) + I(t) + \sigma\eta_i(t) . \qquad (5.2)$$

Here uniform all-to-all coupling is assumed, β_e and β_s denote electrotonic and synaptic coupling strengths, and the 'alpha function' describes the post-synaptic excitation at neuron i after neuron j fires at time(s) $t_{jk} < t$: $A(t - t_{jk}) = [(t - t_{jk} - t_d)/\tau_A] \cdot \exp(-(t - t_{jk} - t_d)/\tau_A)$, where $\tau_A = 30$ msec is the synaptic time constant and $t_d = 25$ msec is an estimate of the alpha 2 NE receptor-mediated synaptic delay [180, Fig. 1]. The white noise term $\sigma\eta_i(t)$ represents unmodeled 'fast' synaptic inputs.

Fig. 5.2 shows the orbit in (V_i, q_i)-space of an isolated Rose-Hindmarsh neuron with I_i^b set to produce periodic spiking, and subject to a tonic stimulus $I(t)$ of greater strength than those employed below, superposed on the unperturbed trajectory ($I_i^{ext} \equiv 0$). Like most conductance-based neural models in repetitive firing modes, (5.1) possesses a strongly attracting, normally hyperbolic limit cycle [82], implying that in the presence of moderate perturbations due to coupling and input currents, solutions remain confined to a small neighborhood of the original orbit. This permits reduction of (5.1) to phase variables, as in Chapter 2, by defining nonlinear polar coordinates and projecting along isochrons [81, 183] onto the unperturbed limit
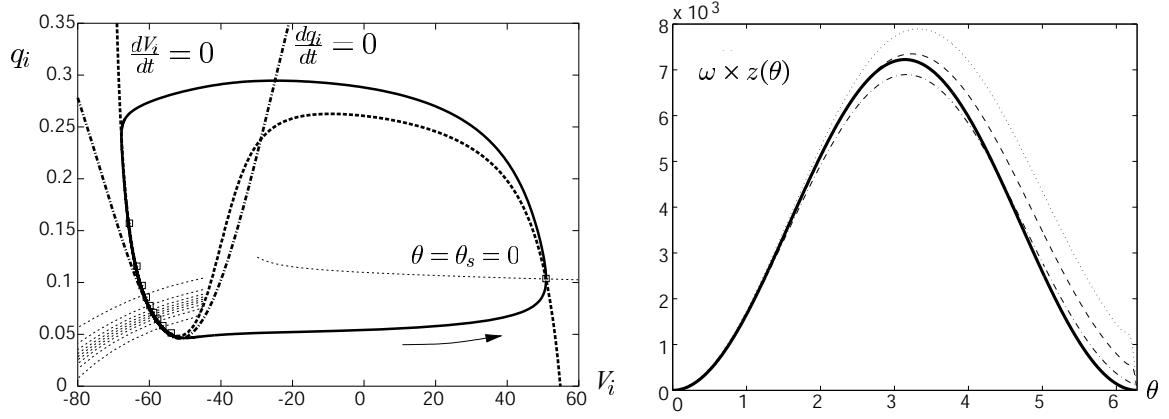
Figure 5.2: (Left) Phase space structure for a repetitively spiking Rose-Hindmarsh neuron (5.1), $I_i^b = 5\mu A/cm^2$. Attracting limit cycle for $I_i^{ext} = 0$ shown solid. Initial conditions on a given isochron (shown dashed) asymptotically approach the same point on the limit cycle as $t \to \infty$; isochrons are equally spaced in phase by $2\pi/10$, with $\theta = 0$ at action potential peak. The thick dashed and dash-dotted lines are nullclines for $\dot{V}_i = 0$ and $\dot{q}_i = 0$, respectively, and squares show points on perturbed limit cycle, equally spaced in time, under tonic stimulus of $I_i^{ext} = 1 \ \mu A/cm^2$. (Right) Reproduced from Fig. 5.3. PRCs for the Rose-Hindmarsh model (5.1) at frequencies $\omega/2\pi \approx 5$ Hz (dotted), $\omega/2\pi \approx 3.2$ Hz (dashed), $\omega/2\pi \approx 1.6$ Hz (dot-dashed). PRCs plotted as $\omega \times z(\theta)$ vs. θ to illustrate that $z(\theta) = \frac{c}{\omega}[1 - \cos\theta]$ with $c = 0.0036 \, (\text{mV} \cdot \text{msec})^{-1}$ (solid) provides an acceptable fit, improving as ω decreases.

cycle (Fig. 5.2):

$$
d\theta_i = \left[\omega_i + z(\theta_i) \left(I(t) + \frac{\beta_e}{N} \sum_{j=1}^{N} \left(V(\theta_j) - V(\theta_i) \right) + \right. \right.
$$

$$
\left. \left. \frac{\beta_s}{N} \sum_{j=1}^{N} \sum_{k} A(t - t_{jk}) \left(V_K - V(\theta_i) \right) \right) + \frac{\sigma^2}{2} z(\theta_i) z'(\theta_i) \right] dt + \sigma z(\theta_i) dW_i(t) \ .
$$

$$(5.3)$$

Here the $\sigma W_i(t)$ are independent Wiener processes with variance $\sigma^2 t$, and ω_i is the frequency of the ith LC neuron, which may vary slowly, e.g. via I_i^b, but is assumed constant over each experimental trial (see below). The phase θ is defined to increase at a constant rate ω_i in the absence of coupling and external inputs, with voltage peak (spike) at $\theta = 0$. As in Chapter 2, the phase response curve (PRC) $z(\theta_i)$ [183, 163, 50], encoding the phase shift due to instantaneous perturbations, multiplies the stimulus and the external noise term; as in Section 3.5 the $\mathcal{O}(\sigma^2)$ term is the 'Ito correction' resulting from changing variables from the stochastic differential equation (5.2) [65]. The functions $V(\theta_i)$ in (5.4) are computed from the unperturbed voltage profile as $V(\theta_i) = V(\omega_i t)$.

As I_i^b increases, Type I neurons undergo a transition from excitability, with a

stable hyperpolarised rest point, to repetitive spiking via a saddle-node bifurcation on a limit cycle [82] (i.e., the SNIPER bifurcation of Chapter 2). Normal form theory was used in Chapter 2 to derive the PRC approximation $z(\theta) = C(\omega)[1 - \cos(\theta)]$ near the bifurcation point [50, 23]. This approximation is reasonable in the frequency range of interest (1-5 Hz); moreover, $C(\omega) = c/\omega$ where $c = 0.0036\,(\text{mV} \cdot \text{msec})^{-1}$: see Fig. 5.2. (As described in Chapter 2, other neuron models yield different PRC forms and ω-scaling, cf. [50].) In particular, the two-compartment LC neuron of [1] yields PRCs similar to those of Fig. 5.2, although the lack of an explicit form for $z(\theta)$ precludes analysis of the type done below.

5.4.2   Coupling effects and synchrony

We now ask how the baseline synchrony in LC neurons, as demonstrated in the experimental data of Fig.5.8 (left), can arise from the synaptic and electrotonic coupling terms in Eqns. (5.2) and (5.4). We use the results of Chapter 4 to determine the firing patterns that the coupling terms will evoke; since we are addressing coherence resulting from coupling (as opposed to phase resetting connected with stimulus onset), we consider (5.4) with $I(t) \equiv 0$. After and applying (stochastic) averaging as in Chapter 3, (5.4) becomes

$$d\theta_i = \left[\omega_i + \frac{1}{N}\sum_{j=1}^{N}(\alpha_e f_e(\theta_j - \theta_i) + \alpha_s f_s(\theta_j - \theta_i)) + z(\theta_i)I(t)\right]dt + \sigma\hat{z}dW_i(t)\ ,\ \ (5.4)$$

where $\sigma\hat{z}_\omega = \left(\frac{1}{2\pi}\int_0^{2\pi}\sigma^2 z(\theta)^2 d\theta\right)^{1/2} = \sqrt{\frac{3}{2}}\frac{\sigma c}{\omega}$ (recall Chapter 3, Section 3.5). Here f_e and f_s are the averaged electrotonic and synaptic coupling terms, shown in Fig. 5.3

Note that *both* coupling functions $f_e(\theta) \approx \sin(\theta)$, and $f_s(\theta) \approx \sin(\theta) + k_s$, where $k_s \approx -0.2$. When $\omega_i \equiv \omega$, it follows from Proposition 4.3.5 that among all of the rotating block and two-block solutions, *only* the in-phase (i.e., perfectly synchronized) periodic orbit is asymptotically stable for any combination of (weak) electrotonic and synaptic coupling and also that, if the coefficients of these terms are positive (as they must be), that the in-phase state is globally attracting in the absence of noise.

The correlation data of [167, Fig. 4A] indicates only *partial* synchrony, even in the phasic mode. We find below that (as for the HH equations, Fig. 4.10(b)) this can be captured by coupling terms that are weak compared with the stimulus, frequency heterogeneity, and noise: cf. Fig. 5.8. Neglecting such weak coupling, any given cell is approximately governed by:

$$
\begin{aligned}
d\theta &= \left[\omega + z(\theta)I(t) + \frac{\sigma^2}{2}z(\theta)z'(\theta)\right]dt + \sigma z(\theta)dW(t) \\
&\overset{\triangle}{=} v(\theta, t)dt + \sigma z(\theta)dW(t)\ .
\end{aligned}
\qquad (5.5)
$$

Here and henceforth we drop the subscripts i and let θ and ω represent the phase and frequency of a typical neuron. Via comparison with simulations of the fully coupled biophysical model (5.1-5.2), we demonstrate below that this greatly reduced equation
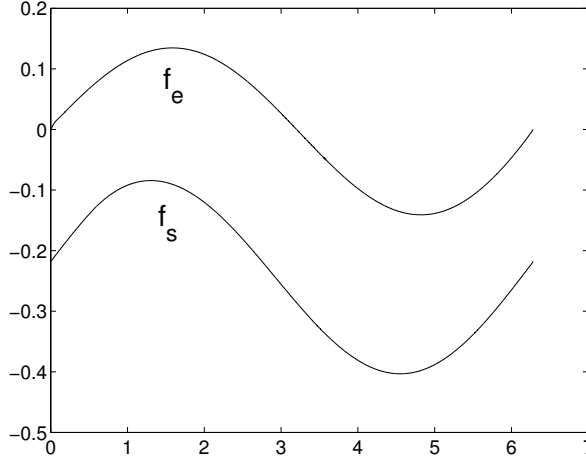
Figure 5.3: Coupling functions for RH neurons, computed for frequency $\omega/2\pi \approx 3.5$ Hz, synaptic delay $t_d = 25$ msec and time constant $\tau_A = 30$ msec; f_e varies little with frequency, f_s exhibits stronger dependence on ω due to time-dependence in synaptic transmission.

provides an adequate model (cf. Fig. 5.9).

5.4.3 Modeling LC modes, frequency variability, and stimuli

To further develop our model, we review the data analysis leading to Fig. 5.1. Target detection PSTHs were obtained by averaging single-cell recordings over one session (≈ 100 trials), after separating epochs of good and poor behavioral performance according to error rates. These epochs correspond to phasic vs. tonic LC modes, respectively [167, Fig. 1A]. Eriksen PSTHs derive from single- or multi-unit recordings, and multiple sessions. No clear tonic episodes were identified in the Eriksen data, although significant frequency variations were seen in individual cells over time, and among multiple cells at any given time; see Fig. 5.4. To reproduce the experimental data, frequencies ω will be drawn from appropriate narrow (for target detection) and broad (for Eriksen) distributions, noise variances fitted to match interspike interval distributions, and coupling strengths chosen to approximately reproduce correlograms.

Since decisions take longer in more complex tasks, we assume that LC inputs due to stimuli are briefer and more intense in target identification than in the Eriksen task. We take a simple square wave input of intensity $\bar{I}_{TD}$, onset time t_1, and offset time $t_2 = t_1 + d_{TD}$ (i.e., of duration d_{TD}) in the former case, and in the latter, a function that rises exponentially towards $\bar{I}_E$ for a period d_E and decays exponentially thereafter. Moreover, the Eriksen data does not indicate performance-dependent variations in baseline LC activity, and incorrect PSTHs keyed on response (rather than stimulus, as in Fig. 5.1) have peak activities similar to corrects, but it does display significantly broader reaction time distributions [33]. We therefore ascribe differences between
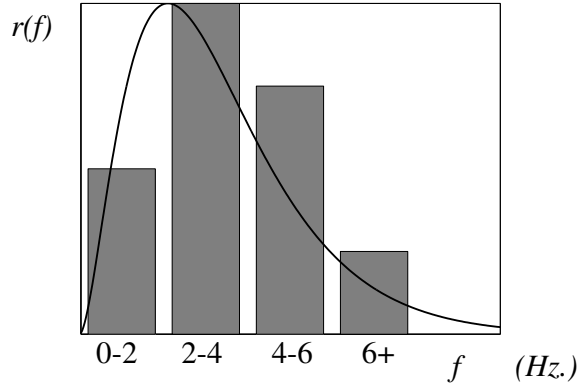
Figure 5.4: Bars: Estimated distribution of LC spiking frequencies at a fixed time. Solid line: Gamma distribution fit: $\Gamma(f) = \frac{1}{\Gamma(\beta)\phi^\beta} f^{\beta-1} \exp(-\frac{f}{\phi})$, with $\phi = 1.2, \beta = 2.75$ chosen to match mean and minimize least-squares difference within each quartile.

the correct and incorrect PSTHs of Fig. 5.1 to variable input latencies in this more complex task. These and all other parameter choices are specified in Section 5.6.1, below.

5.5   Probabilistic analysis

We now apply the results of Chapter 3 on response dynamics of neural oscillator populations to characterize how LC firing patterns depend on both baseline rates and on the duration of LC inputs associated with task stimuli.

   As discussed above, to describe the LC we adopt the HR reduction of the classical Connor model; hence, neglecting noise for now, the results of Chapter 3 for the SNIPER bifurcation (for which the HR model is our archetype) are applicable to the LC model. The right hand panels of Fig. 5.5 show $FL(t)$ for two different inputs. The short, strong (target detection) inputs yield post-peak intervals of depressed firing and substantial 'ringing,' while the protracted input gives less ringing. We also show histograms computed via direct numerical simulations of the Rose-Hindmarsh Eqns. (5.1), indicating that (as in Chapter 3), apart from a slight time stretch due to the PRC approximation, the reduction to a phase equation is remarkably accurate. We now collect those results which will yield insights into these important aspects of LC dynamics.

   First, under the same assumption that PSTHs are averaged over uniform phase probabilities for individual LC neurons, LC firing rates are described by

$$FL(t) = \omega\rho(0,t) = \frac{\omega}{2\pi} \left[\frac{\omega + \bar{I}\left(1 - \frac{c}{\omega}\cos(\Theta_{0,t}(t_1))\right)}{\omega + \bar{I}\left(1 - \frac{c}{\omega}\cos(\Theta_{0,t}(\tilde{t}_2))\right)} \right] . \tag{5.6}$$

This follows from inserting the PRC for the SNIPER bifurcation $z(\theta) = \frac{c}{\omega}(1 - \cos(\theta))$ (Eqn. (2.36)) into Eqn. (3.8) and taking the limit $\psi \to \theta_s$ in that equation (see

Section 3.3.2). After performing an integral, we have the explicit formulas

$$\Theta_{\theta,t}(\tilde{t}_2) = \theta - \omega(t - \tilde{t}_2) \,, \tag{5.7}$$

$$\Theta_{\theta,t}(t_1) = 2\arctan\left\{\sqrt{\frac{\omega}{b}}\tan\left[\arctan\left(\sqrt{\frac{b}{\omega}}\tan\left[\frac{\Theta_{\theta,t}(\tilde{t}_2)}{2}\right]\right) - \frac{1}{2}(\tilde{t}_2 - t_1)\sqrt{\omega b}\right]\right\} \,, \tag{5.8}$$

where $b = \omega + 2c\bar{I}/\omega$. (Recall that the stimulus switches on at time t_1 and off at t_2; for $t < t_1$, we assume the uniform distribution $\rho(0, t) \equiv 1/2\pi$.) This expression shows that the critical stimulus duration yielding no post-stimulus response is $P = \frac{2\pi}{\sqrt{2c\bar{I}+\omega^2}}$ for our model LC neurons; additionally equipped with (3.10) from Chapter 3, we conclude that $\theta_{max} = \theta_{min} = \pi$ and hence that the stimulus durations giving maximal and minimal post-stimulus response are $d_{max} = d_{min} = nP + P/2$.

To compare responses for different values of ω, $\bar{I}$ and d, we define the peak and refractory indices $R_p(d)$ and $R_r(d)$ (dependent on stimulus duration $d = t_2 - t_1$) as

$$R_p(d) = \frac{FL_{max}(d) - FL_{base}}{FL_{base}} \;\;; \qquad R_r(d) = \frac{FL_{base} - FL_{min}(d)}{FL_{base}} \,, \tag{5.9}$$

where the baseline value is $FL_{base} = \omega/2\pi$. Fig. 5.6 (right) illustrates the d-dependence of R_p and R_r; in particular, from the calculations summarized in Tables 3.2 and 3.5 we have:

$$R_p^{max} = R_p(nP + P/2) = \frac{2c\bar{I}}{\omega^2} \;\;; \qquad R_r^{max} = R_r(nP + P/2) = \frac{2c\bar{I}}{2c\bar{I} + \omega^2} \,, \tag{5.10}$$

as shown in Fig. 5.6 (left and center). Note that R_p^{max} is proportional to stimulus strength over frequency squared, which quantifies our first main result: **Maximum LC response is elevated in populations with baseline slower firing rates.** This effect, which in our model primarily derives from the c/ω factor in the PRC, is clear in the bottom left and top right panels of Fig. 5.5; note that (5.10) implies $FL_{max} - FL_{base} \sim 1/\omega$.

As Fig. 5.6 (right) demonstrates, if inputs due to stimuli are sufficiently short compared with the (ω-dependent) response period:

$$d < P(\omega) = \frac{2\pi}{\sqrt{2c\bar{I} + \omega^2}} \,, \tag{5.11}$$

then $\rho(\theta, t_2, \omega)$ necessarily exhibits a peak and a trough, so that successive episodes of enhanced and depressed spiking ensue following stimulus offset. Longer inputs may or may not have this effect: they can end near 'integer points' $d \approx nP$, leaving $\rho(\theta, t_2, \omega) \approx 1/2\pi$, or at $d \approx nP + P/2$, leaving stronger post-stimulus effects; cf. Fig. 5.5. Furthermore, during the stimulus itself, firing rates do not dip below their baseline, as per point **(V)** and Table 3.3 of Chapter 3 and as demonstrated in Fig. 5.5 (d).

Equation (3.45) of Section 3.5 expresses the fact that **noise and frequency**
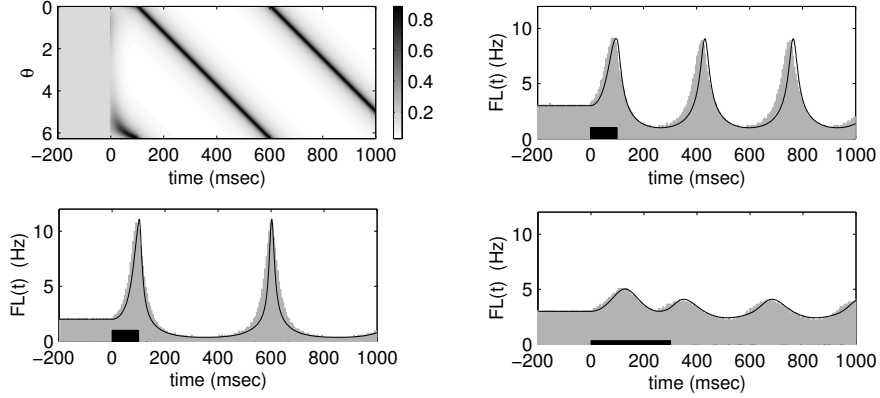
Figure 5.5: (Top and bottom left) Phase density $\rho(\theta, t)$ and flux $FL(t)$ computed from Eqns. (3.6) and (3.8) with $\omega/2\pi = 2$ Hz, $\bar{I} = 0.1$ μA/cm^2, $d = 100$ msec. (Top and bottom right) Fluxes $FL(t)$ for $\omega/2\pi = 3$ Hz, $\bar{I} = 0.1$ μA/cm^2, $d = 100$ msec (top) and $\bar{I} = 0.0333$ μA/cm^2, $d = 300$ msec (bottom). Stimuli indicated by black bars. The 'charge' $\bar{I}d = 10$ μA $\cdot$ msec/cm^2 in both cases. Gray bars show spike rates computed directly from Rose-Hindmarsh Eqns. (5.1).

heterogeneity cause exponential or faster decay of firing rates to baseline levels. For example, typical variations in $P(\omega)$ for the broad distribution of Fig. 5.4 range from 145 to 205 msec, leading to significantly differing $\rho(\theta, t_2, \omega)$'s, and differing propagation speeds. However, for tight distributions $r(\omega)$, $P(\omega)$ varies little and $\rho(\theta, t_2, \omega)$ travel at approximately the same speed, so the leading peak and depression can be expected to survive averaging over mild oscillator heterogeneity. This leads to the finding: **In systems with narrow frequency distributions, short inputs necessarily lead to intervals of depressed firing following enhanced spiking and stimulus offset.**

This effect is further magnified if we normalise to maintain fixed 'synaptic charge' $\bar{I}d = S$. Now $\bar{I} \propto 1/d$ and (5.10) (Fig. 5.6) shows that brief inputs are yet further enhanced over longer, more diffuse ones. In this case, eliminating $\bar{I}$ from (5.11) yields an explicit input duration for maximal effect:

$$d \approx \frac{P}{2} = \frac{1}{\omega^2} \left(\sqrt{c^2 S^2 + \pi^2 \omega^2} - cS \right) . \tag{5.12}$$

5.6 Comparison with experimental data

5.6.1 Parameter fitting

To compare model predictions with data, we first determine appropriate frequency distributions $r(\omega)$ for *single* neurons recorded over long durations, characterized by mean μ^ω and variance γ^ω, and r.m.s. noise strength σ, by seeking parameter values for which model realizations match both an empirical interspike interval (ISI) histogram
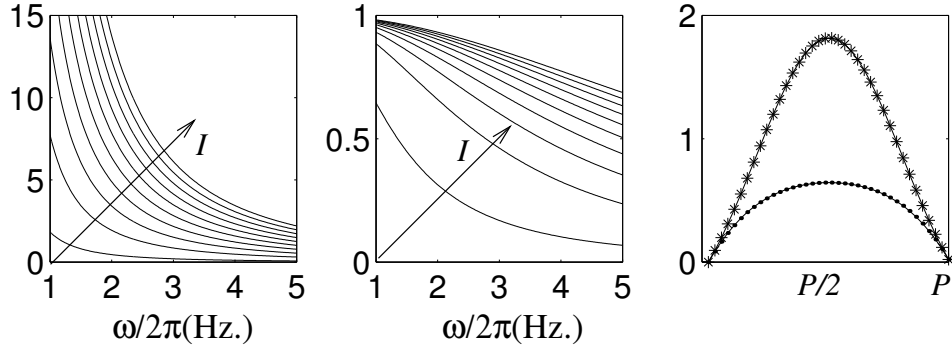
Figure 5.6: (Left, center) R_p^{max} and R_r^{max} as functions of ω for values of $\bar{I}$ evenly spaced between 0.01 and 0.30 μA/cm^2. (Right) $R_p(d)$ (stars) and $R_r(d)$ (dots) for $\omega/2\pi = 3.2$ Hz, $\bar{I} = 0.1$ μA/cm^2 and d ranging between 0 and P. The greatest post-stimulus ringing of the firing rate will occur for values of d around $P/2$ (where $R_{p,r}(d)$ are largest).

and correlations between neighboring ISIs:

$$r_1 \triangleq \frac{\mathbb{E}\left\{(y_j - m)(y_{j+1} - m)\right\}}{\mathbb{E}\left\{(y_j - m)^2\right\}} \ . \tag{5.13}$$

Here subsequent ISIs are labeled y_j and $\mathbb{E}$ denotes expectation, $m = \mathbb{E}\{y_j\}$. The process $\{y_j\}$ is assumed stationary so r_1 and m are independent of j. Variability is assumed due to: 1) slow drift in baseline frequency ω, and 2) rapid input current fluctuations modeled through the $\sigma z(\theta)dW(t)$ term in (5.4). Thus $y_j = y_j^\omega + \eta_j$, where y_j^ω are the noise-free drift values, and η_j causes additional variance $(\gamma^\sigma)^2$ due to rapid noise. If the drift is sufficiently slow then $\mathbb{E}\{(y_j - m)(y_{j+1} - m)\} \approx \mathbb{E}\{(y_j^\omega - m)^2\} \triangleq (\gamma^\omega)^2$, and (5.13) becomes

$$r_1 \approx \frac{(\gamma^\omega)^2}{(\gamma^\sigma)^2 + (\gamma^\omega)^2} = 0.1 \ , \tag{5.14}$$

where we appeal to independence of the η_j and insert the numerical value derived from the data of Fig. 5.7, recorded from a single LC neuron in an Eriksen session.

Eqn. (5.14) constrains the ratio of slow $((\gamma^\omega)^2)$ to fast $((\gamma^\sigma)^2)$ ISI variances, the breadth of the ISI histogram constrains the magnitude of these variances, and the mean frequency μ^ω may be estimated directly from the ISI mean m. Guided by this and by analytical expressions for, e.g., barrier hitting times relating σ and γ^σ, Monte-Carlo simulations suggest a Gaussian distribution $r(\omega)$ with mean 1.69 Hz and standard deviation 0.47 Hz, and r.m.s. noise strength $\sigma = 0.45$. This yields the model ISI distribution of Fig. 5.7. To match the baseline data for the single neuron PSTHs of the target detection task, we rescale the center frequencies to 2 and 3 Hz respectively for the phasic and tonic modes, while keeping the ratio of mean to standard deviation constant. For the multi-neuron Eriksen data, we use the broader Gamma distribution of Fig. 5.4. (Recall that these different frequency distributions
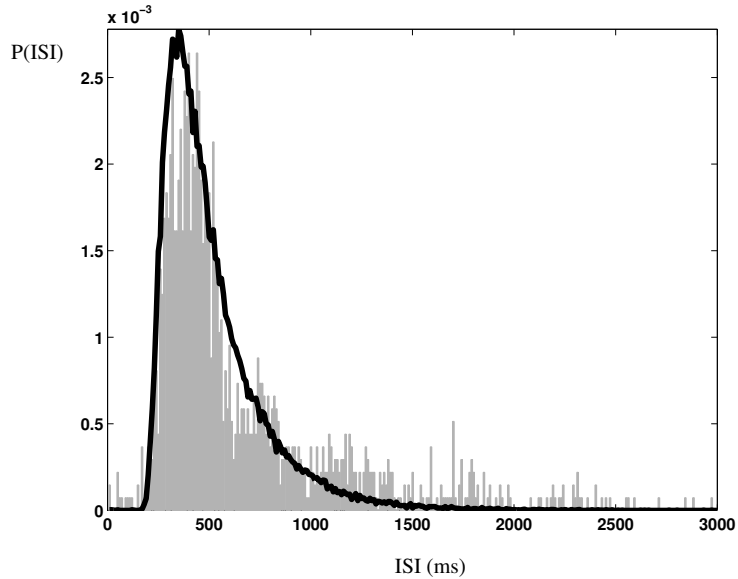
Figure 5.7: Empirical (gray bars) and model (solid line) histograms of baseline ISIs for a single neuron during the Eriksen task. For empirical data, windows of 1.5 sec. following each stimulus are removed to isolate baseline frequencies. Histogram bins are 10 ms wide.

are realized by different distributions of baseline currents I_i^b, see Section 5.7.) We maintain $\sigma = 0.45$ throughout.

Synaptic and electrotonic coupling strengths were chosen to qualitatively capture the experimental cross-correlograms for phasic and tonic episodes given in [167]; see Fig. 5.8. As described in Section 5.3.2, [167] reports that baseline firing rates (among the entire LC population) are not only slower than in the tonic mode but are also more tightly distributed. Thus, in the phasic (tonic) mode, we draw spike frequencies from a Gamma distribution with $\beta = 3, \phi = 0.667$ ($\beta = 3, \phi = 1$), giving mean 2 Hz (3 Hz) and standard deviation 1.16 Hz (1.73 Hz); cf. Fig. 5.4. (However, as described in the previous paragraph, only some of these frequencies contribute to the single-neuron PSTHs of the target detection task.) We then require that a central subgroup of oscillators are largely synchronous (asynchronous). This yields $\beta_s = 0.01$, $\beta_e = 0.05$. Note that, unlike [167], we take the *same* coupling strengths for phasic and tonic modes, showing that increased synchrony can result solely from a tighter distribution of phasic frequencies.

Finally, we found via interactive simulations the following appropriate inputs $I(t)$. For the target detection task we found that a square wave of height $\bar{I} = 0.125 \ \mu\text{A}/\text{cm}^2$ and duration $d = 110$ msec was satisfactory. Since the Eriksen data is averaged over all conditions (congruent and incongruent stimuli) and presumably involves more complex cognitive processing, a more diffuse input is appropriate. We adopted an exponentially rising and falling function, with rise duration $d = 180$ msec and rise and fall time constants 75 and 90 msec respectively and maximum height $\bar{I} = 0.22$
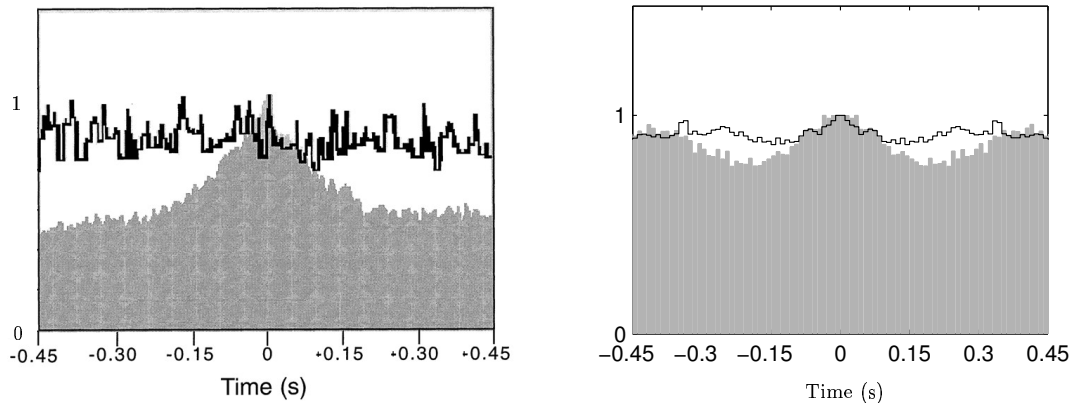
Figure 5.8: Normalized cross correlograms for phasic LC mode (filled histogram) and tonic mode (solid line). (Left) from [167, Fig. 4] for two simultaneously recorded LC neurons. (Right) model results derived from mean ± 1 standard deviation of 100 oscillator population. In both cases, central peak indicates increased synchrony in phasic mode.

μA/cm^2. Moreover, reaction time (RT) distributions have significantly greater standard deviation than for target detection: 114 and 241 msec for correct and incorrect respectively [33], compared to $\approx$ 34 and 53 msec for phasic and tonic modes respectively [167]. We therefore averaged over Gaussian distributions of onset times with standard deviations of 38 and 80 msec in the Eriksen task, assuming that variability in input arrival times at LC contributes about one third of total RT variability. Because our simulations indicate that the much smaller RT variability in the target detection task produces only minor effects (see below), we used fixed latencies in modeling this task. In all cases, since LC input lags visual stimulus, we include a time delay of 90 msec [10]. Thus, as noted in Section 5.3.2, the results of Section 5.5 *predict* that the target detection and Eriksen inputs must differ qualitatively.

5.6.2   Comparison of model and empirical PSTH data

Fig. 5.9 shows model PSTH data for the target detection and Eriksen tasks, obtained in three ways: 1) by numerical solution of (3.1) in the presence of noise ($\sigma \neq 0$), followed by averaging over the frequency distributions derived above; 2) via direct simulations of a set of $N = 100$ globally-coupled Rose-Hindmarsh equations (5.1) representative of the same distributions, excited by independent Brownian noise currents of appropriate strength; and 3) directly from the noise-free expression (5.6) averaged over the same frequency distributions. The probabilistic effects considered above are clear: population averaging and noise combine to damp the periodic ringing of the noise-free single-frequency data of Fig. 5.5 (cf. the decay rate bound (3.45)).

These results confirm that reduction to a single phase equation (5.5) and the
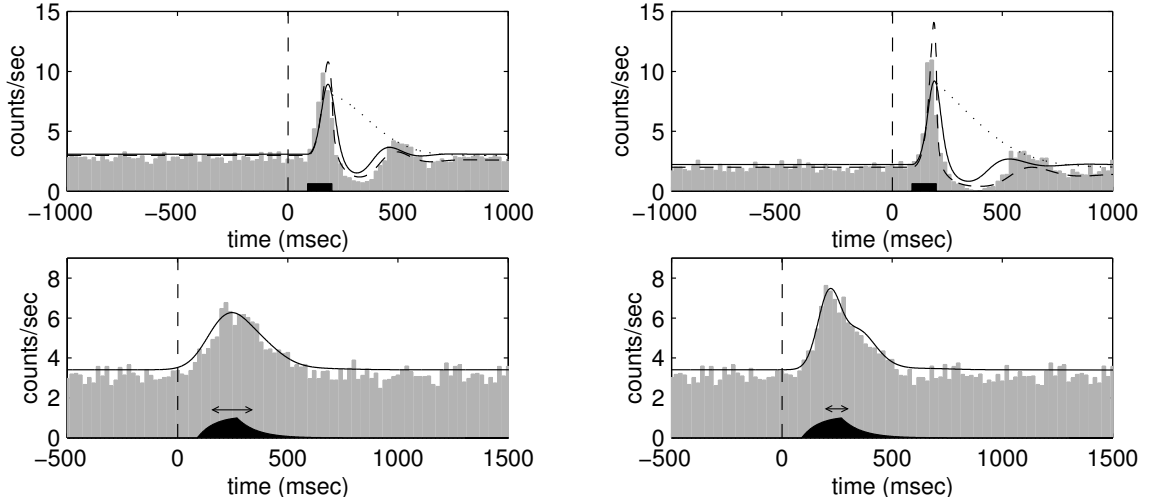
Figure 5.9: Model PSTHs computed from solution of Eqn (3.1) (solid) and from Eqn. (5.6) (dashed), averaged over neuron frequency distributions and with stimuli (shown as filled black; arrows above stimuli for Eriksen task indicate variability in stimulus onset) and all other parameters as described in text. Gray bars show results of simulating 100 Rose-Hindmarsh neurons for multiple trials. Decay bound of (3.45) shown dotted. Top row: target detection task for poor performance/tonic mode (left) and good performance/phasic mode (right); bottom row: Eriksen task for incorrect (left) and correct (right) responses. Compare with the experimental PSTHs of Figure 5.1 above.

probabilistic theory developed above provide good descriptions of the coupled Rose-Hindmarsh system, and that the decay rate bounds of Section 3.5 are reasonable. The noise-free limit (5.6) is a useful qualitative estimator of PSTHs, although target detection phasic/tonic response ratios are significantly less than 9/4 predicted by (5.10), due to the high noise level that selectively damps the sharply peaked phase densities arising at low frequencies.

The model results of Fig. 5.9 qualitatively reproduce the PSTHs of Fig. 5.1, with the major quantitative discrepancy that enhancement of response magnitude for phasic relative to tonic states in target detection captures only a part of that reported in [167]. In terms of the measure R_{mag} that characterizes enhanced spiking following stimulus [10], our model predicts a ratio $R_{mag}(\text{phasic})/R_{mag}(\text{tonic}) \approx 1.3$, compared to the value 3.4 of [167]. Hence, additional mechanisms, beyond the frequency effects studied here, must be operative in the phasic/tonic transition. For example, noise levels σ may be elevated in the tonic mode (in addition to mean current values I_b). Our simulations (not reported here) confirm that this increases the relative magnitude of phasic mode responses, as predicted in [86]. Averaging over a slightly broader distribution of input onsets in the tonic mode than in the phasic mode, as suggested by reaction time distributions in the two tasks, further enhances phasic vs. tonic responses, although the small RT variance shows that this is a minor effect in target

detection. Electrotonic coupling changes may also play a role as in [167]. We note that recent additional analyses of target detection data, in which recordings were grouped by baseline rate without reference to tonic and phasic behavioral modes, revealed small differences in R_{mag} similar to those reported here.

Our model reveals that the tonic/phasic frequency difference contributes to the variation between PSTHs for poor and good target identification performance, while in the Eriksen data, for which baseline frequencies are similar, PSTH differences can be accounted for by variations in stimulus arrival times originating in earlier processing. Moreover, diffuse stimuli in the latter case eliminate the depressed post-activation spiking seen in target identification.

5.7   Discussion

We have shown that a biophysical model of coupled LC neurons can be reduced to a stochastic differential equation for the phase of a given cell, and that a probabilistic formulation and averaging over suitable frequency distributions allows one to model and analyze peri-stimulus time histograms derived from single and multi-cell LC recordings. Our model supplements that of [167], and our analysis reveals explicit parameter dependencies, including the effects of stimuli appropriate to two different cognitive tasks.

In [167], electrotonic coupling variations were proposed as the cause for transitions between tonic and phasic LC modes, and hence for differences in PSTHs associated with poor and good target detection performance. In the model presented here, while coupling clearly affects synchrony, the key factor influencing PSTHs averaged over many trials is the LC spike rate, governed by the baseline currents I_i^b. In any case, our baseline rate explanation differs from the electrotonic coupling mechanism of [167]; also, in that paper the I_i^b were set in the excitable range, so that noise and other external inputs were *necessary* for spiking. From simulations of subthreshold networks of coupled Rose-Hindmarsh neurons with noise-driven firing at $2-3$ Hz., we found that reproducing post-stimulus periods of depressed activity requires strong collateral coupling among LC neurons in both phasic and tonic LC modes. The same conclusion held for solutions of the corresponding (coupled) phase density equations derived from the full 'theta model' [50] (not reported here). In this high noise, high coupling regime, different mechanisms for the phasic to tonic transition may dominate.

Since we assume here that frequencies are distributed more tightly in the slower phasic mode, we obtain enhanced phasic mode synchrony *without* changing coupling strength: Fig. 5.8; this differs from the subtler mechanism of [1]. In sum, we see synchrony as a *correlate* of elevated LC response, rather than its primary cause. For the LC *in vivo*, the synchronizing effects identified here and in [167, 1] may all be relevant. Additional effects of stronger coupling terms, noise, and subthreshold neurons may also be important and are under investigation.

Possible explanations for decreased I_i^b include reduced inputs from other neurons afferent to the LC. The anterior cingulate cortex (ACC), a prefrontal area involved in cognitive control, has recently been shown to have excitatory (presumably gluta-

matergic) projections to the LC [138, 99, 100]. Our findings suggest that the ACC may send decreased excitation to the LC in the phasic vs. tonic mode. Intriguingly, small inhibitory (GABA-ergic) neurons have been found among in a peri-LC region and are known to project to LC neurons and dendrites [11]. Several areas, including the prefrontal cortex, innervate this region, suggesting a pathway by which input currents I^b, and hence baseline firing, may be regulated by increased inhibition in the phasic mode.

Decreased firing rates in the phasic vs. tonic mode could also result from neuromodulators. For example, in some cases direct application of the neuropeptide corticotropin releasing factor (CRF) increases LC baseline activity and simultaneously decreases responses to sensory stimuli [170]. It has also been found that that the alpha2 adrenoceptor agonist clonidine (or ST-91) can decrease baseline activity and increase response [7]; the neuromodulator corresponding to this drug is norepinephrine, which presumably could be sent to the LC from other noradrenergic brain areas. Many other examples of such 'modulatory' effects of neurotransmitters or exogenous inputs exist for neurons in other brain areas [8]. Finally, we note that since synaptic coupling among LC neurons is inhibitory, it transiently reduces net input currents, thus effectively decreasing I_i^b if LC neurons are sufficiently decorrelated.

The present analysis therefore provides a simple explanation for how mechanisms which set baseline firing rate, such as background levels of exogenous input, neuromodulators, or pharmacological agents will influence the response of a neural population to pulsed stimuli (cf. [86, 60]): indeed, it shows that this dual effect on baseline rates and evoked response is intrinsic to the dynamics of neural groups. The LC phasic and tonic modes are an example, but the dual effect occurs in numerous other brain areas and neurons [8]. This intrinsic baseline rate mechanism joins a list of others: in addition to altered electrotonic coupling [167], other mechanisms for simultaneous effects on baseline and stimulus-evoked firing have been proposed, including simultaneous transmitter actions at multiple receptors [10], alterations in specific second messenger pathways and ion conductances [123, 62].

In addition to the predictions regarding *in vivo* baseline LC inputs just described, our analysis also provides a prediction about inputs evoked by task-related stimuli. That is, neurons that project to the LC and evoke responses should remain active longer following stimuli in complex tasks such as the Eriksen paradigm than in simpler ones like target detection. This is consistent with the notion introduced above (assumptions (A3) and (A4) of Sect. 5.3.2) that the LC is driven by accumulating activity in decision areas, as this activity may be expected to accumulate more gradually in complex decision tasks.

In recent related work [168, 70] abstracted models of LC population activity that modify gains in connectionist networks have been shown to capture neuromodulatory effects on cognitive performance. Before turning to investigate these gain effects in detail in Part II of this dissertation, we remark that the present LC model, derived from the neural substrate, offers simplification comparable to [168, 70] as well as suggesting, in the coupled multi-unit phase model of Eqn. (5.4), a middle ground between those abstractions and the full Rose-Hindmarsh system of (5.1-5.2).

Appendix to Part I

Equations for the neural models

The Rose-Hindmarsh equations:

$$
\begin{aligned}
\dot{V} &= [I^b - g_{Na}m_\infty(V)^3(-3(q - Bb_\infty(V)) + 0.85)(V - V_{Na}) \\
&\quad - g_K q(V - V_K) - g_L(V - V_L)]/C \\
\dot{q} &= (q_\infty(V) - q)/\tau_q(V) \\
q_\infty(V) &= n_\infty(V)^4 + Bb_\infty(V) \,, \quad b_\infty(V) = (1/(1 + \exp(\gamma_b(V + 53.3))))^4 \,, \\
m_\infty(V) &= \alpha_m(V)/(\alpha_m(V) + \beta_m(V)) \,, \quad n_\infty(V) = \alpha_n(V)/(\alpha_n(V) + \beta_n(V)) \,, \\
\tau_q(V) &= (\tau_b(V) + \tau_n(V))/2 \,, \quad \tau_n(V) = T_n/(\alpha_n(V) + \beta_n(V)) \,, \\
\tau_b(V) &= T_b(1.24 + 2.678/(1 + \exp((V + 50)/16.027))) \,, \\
\alpha_n(V) &= 0.01(V + 45.7)/(1 - \exp(-(V + 45.7)/10)) \,, \\
\alpha_m(V) &= 0.1(V + 29.7)/(1 - \exp(-(V + 29.7)/10)) \,, \\
\beta_n(V) &= 0.125 \exp(-(V + 55.7)/80) \,, \quad \beta_m(V) = 4 \exp(-(V + 54.7)/18) \,.
\end{aligned}
$$

$V_{Na} = 55$ mV , $V_K = -72$ mV , $V_L = -17$ mV , $g_{Na} = 120$ mS/cm^2 ,
$g_K = 20$ mS/cm^2 , $g_L = 0.3$ mS/cm^2 , $g_A = 47.7$ mS/cm^2 ,
$C = 1$ μF/cm^2 , $I_i^b = 5$ μA/cm^2 , $\gamma_b = 0.069$ mV^{-1} ,
$T_b = 1$ msec , $T_n = 0.52$ msec , $B = 0.21\, g_A/g_K$.

The Fitzhugh-Nagumo equations:

$$
\begin{aligned}
\dot{V} &= [-w - V(V - 1)(V - a) + I^b]/C \\
\dot{w} &= \epsilon(V - g_a w)
\end{aligned}
$$

$g_a = 1$, $\epsilon = 0.05$, $a = 0.1$ mV , $C = 1$ μF/cm^2.

The Hodgkin-Huxley equations:

$$
\begin{aligned}
dV/dt &= 1/C(I - g_{Na}h(V - V_{Na})m^3 - g_K(V - V_K)n^4 - g_L(V - V_L)) \\
dm/dt &= a_m(V)(1 - m) - b_m(V)m \\
dh/dt &= a_h(V)(1 - h) - b_h(V)h \\
dn/dt &= a_n(V)(1 - n) - b_n(V)n \\
a_m(V) &= 0.1(V + 40)/(1 - \exp(-(V + 40)/10)) \\
b_m(V) &= 4\exp(-(V + 65)/18) \\
a_h(V) &= 0.07\exp(-(V + 65)/20) \\
b_h(V) &= 1/(1 + \exp(-(V + 35)/10)) \\
a_n(V) &= 0.01(V + 55)/(1 - \exp(-(V + 55)/10)) \\
b_n(V) &= 0.125\exp(-(V + 65)/80)
\end{aligned}
$$

$$V_{Na} = 50 \text{ mV} , \ Vk = -77 \text{ mV} , \ V_L = -54.4 \text{ mV} , \ g_{Na} = 120 \text{ mS/cm}^2$$
$$g_K = 36 \text{ mS/cm}^2 , \ g_L = .3 \text{ mS/cm}^2 , \ C = 1 \ \mu\text{F/cm}^2$$

The Morris-Lecar equations:

$$
\begin{aligned}
\dot{V} &= [g_{Ca}m_\infty(V)(V_{Ca} - V) + g_K w(V_K - V) + g_L(V_L - V) + I^b]/C \\
\dot{w} &= \phi(w_\infty(V) - w)/\tau_w(V) \\
m_\infty(V) &= 0.5(1 + \tanh((V - V_1)/V_2)) \\
w_\infty(V) &= 0.5(1 + \tanh((V - V_3)/V_4)) \\
\tau_w(V) &= 1/\cosh((V - V_3)/(2V_4))
\end{aligned}
$$

$$\phi = 0.23 , \ g_L = 2 \text{ mS/cm}^2 , \ g_{Ca} = 4 \text{ mS/cm}^2 , \ g_K = 8 \text{ mS/cm}^2 , \ C = 20 \ \mu\text{F/cm}^2$$
$$V_K = -84 \text{ mV} , \ V_L = -60 \text{ mV} , \ V_{Ca} = 120 \text{ mV}$$
$$V_1 = -1.2 \text{ mV} , \ V_2 = 18 \text{ mV} , \ V_3 = 12 \text{ mV} , \ V_4 = 17.4 \text{ mV}$$

PART II: Neural Integrators

Chapter 6

Optimal gain for simple decision models with known input schedules

6.1 Chapter outline

We review simple connectionist and firing rate models for mutually inhibiting pools of neurons that discriminate between pairs of stimuli. Both are two-dimensional nonlinear stochastic ordinary differential equations, differing in how inputs and stimuli enter. A key parameter is gain: the maximum slope of the sigmoidal activation function. We develop piecewise-linear and purely linear models, and one-dimensional reductions to Orstein-Uhlenbeck processes that can be viewed as linear filters. We then pose and solve the optimal gain problem for the reduced models, finding explicit gain schedules that minimize error rates for time-varying stimuli. We relate these to time courses of norepinephrine release in cortical areas, and argue that transient firing rate changes in the brainstem nucleus locus coeruleus may be responsible for approximate gain optimization. The material of this chapter forms a part of [20].

6.2 Introduction

The psychological and neural bases of decision making are active areas of inquiry in cognitive science [152, 73, 151, 72, 156, 122, 152, 161, 113, 139, 140, 169, 145, 184]. There is a wealth of data on simple decision tasks which require discrimination among alternative stimuli as quickly and accurately as possible. Typically, this discriminatory process has been modelled as a competition among different neural populations, each representing alternate interpretations of the current stimulus [38, 169]. Recent direct recordings in visual and motor areas of monkeys performing sensory discrimination tasks support this interpretation by revealing that, following training, certain 'decision' neurons become selective for different stimulus alternatives, and upon presentation of the relevant stimulus their firing rates gradually increase accordingly; when these rates cross thresholds, the corresponding behavioral response is initiated (e.g. [152, 73, 151, 145, 72]). This neural evidence adds to behavioral evidence noted below, suggesting that decisions are made by comparing integrated 'weights of

evidence,' encoded by the firing rates of neural groups. Here, we explore the computational mechanisms required to optimize such a process.

The stimuli relevant to making a decision are often not static: their saliences may change over time. In the simplest case, a change occurs only at the moment when the stimulus itself appears. This is typically modelled in simulations of decision tasks (e.g., in [36, 21, 31], cf. [113]) by dividing the task into two distinct periods: a preparatory period, in which no stimulus is present, and a trial period, in which a stimulus of constant discriminability is presented. Alternatively, stimulus discriminability may change in a stepwise manner or vary continuously during the trial period.

The following specific example motivates our analysis of two specific cases in Section 6.3.4. In the 'moving dots' paradigm of the two alternative forced choice task [18, 156, 72] a display of moving dots is presented, and the subject must indicate whether a majority of dots is moving to the right or the left. In the simplest case, the subject focuses on a neutral fixation point during the preparatory period, after which the dots appear, with a certain 'coherent' fraction moving either left or right, and the rest moving randomly. A variant is obtained by showing a zero coherence display of dots during the preparatory period, and suddenly increasing coherence to a fixed value.

Even if external stimuli have constant strengths, their representations in neural populations that decide between alternative hypotheses may *gradually* rise, due to accumulating activity in input layers, fluctuations in attention, or both [125, 39, 167, 70]. Another possible source of time varying salience is the increasing noise levels that may accompany higher firing rates. A richer situation, in which the stimulus salience increases and decreases over time, is explored in [95]. A focus of the present chapter is how stimuli with time dependent salience can be *optimally* processed in simple neurally-based models of decision networks. We study the reduction of such networks to linearized, one-dimensional approximations (cf. [169, 21, 13]) for which optimality conditions can be fully characterized, and identify two distinct mechanisms, one involving intrinsic properties of decision networks and the other involving external modulation, that can implement optimal processing of time-varying stimuli.

Optimality principles have found wide application in psychology and neuroscience (e.g. [12, 3, 56]). In particular, Stone applied the optimal Sequential Probability Ratio Test (SPRT) to model behavioral data in a two-alternative forced choice task [161]. This was followed by the extensive work of Laming [113]. The SPRT computes time-dependent likelihood ratios between the probabilities of two competing hypotheses, a procedure equivalent to the signal processing strategy that maximizes signal-to-noise ratio in the difference between two incoming stimuli. For stimuli with constant signal-to-noise ratios, the SPRT is equivalent, in an appropriate continuum limit, to the constant-drift diffusion model, which has been shown by Ratcliff and others to fit a wide variety of behavioral data (see [139, 140] and references therein) and also to describe the dynamics of neural firing rates in sensori-motor brain areas [151, 72], cf. [159]. Specifically, in [72], the notion of reward rate is introduced for the constant-drift diffusion model, and [13] shows that higher performing subjects do optimize this quantity in a specific behavioral task. However, although [113] does

110

allow for accumulation of noise to have occurred before stimulus presentation (see Laming's appendix A7), in all these studies the decision process is modelled only *after* presentation of a stimulus having constant signal-to-noise ratio; furthermore, the parameters describing processing of incoming information are not explicitly allowed to vary in time.

In this chapter we show how models of mutually inhibiting neural populations can make nearly optimal decisions about the identity of *time-varying* stimuli. This is accomplished via dynamical adjustments in an *effective* gain parameter for the linearized population dynamics. The gain determines the sensitivity of (equilibrium) population firing rates to changes in averaged input currents to the population, and the word 'effective' is used here because these changes can result either from transient variations in the gain parameter describing this sensitivity or directly from the nonlinearities of neural input-output functions. There is much current research into neural mechanisms for the modulation of gain in neural populations, identifying such factors as levels of norepinephrine [167] and the strength of fluctuations in individual neurons comprising the population (e.g. [30, 2, 26]). In particular, [158] proposes a mechanism in which frequency-current curves of individual neurons adapt to match operating ranges to neural inputs, via intracellular calcium signals. This may be viewed as a biophysical implementation of the earlier 'automatic gain control' (see Eqn. (9) of [80] and references therein), which is implemented via multiplicative 'shunting' terms in neural network models and also keeps neural units in the sensitive regimes of their input-output functions. Gain plays a different role in the present chapter: we identify, for three different models, the distinct time-dependent (effective) gain schedules which implement optimal processing strategies for time-dependent signals. These provide predictions for gain manipulations that diverse neural mechanisms may implement to improve task performance.

The balance of the chapter proceeds as follows. In Section 6.3 we introduce the forced and free response decision tasks, and three types of stochastic differential equation (SDE) models for these tasks. We discuss linearized and one-dimensional reductions of them in two rather general cases. In the following Section 6.4, we compute time dependent values of gain that optimize signal processing in the one-dimensional models. This involves calculating gain functions that enable them to implement the classical signal processing notion of matched filters. Section 6.5 interprets these results in terms of cortical norepinephrine (NE) release mediated by the brainstem nucleus locus coeruleus (LC), showing that LC and NE dynamics indeed appear to approximate optimal time courses. Section 6.6 concludes with a brief discussion.

6.3   Models of decision tasks

6.3.1   Decision tasks: the forced and free response protocols

We consider two distinct tasks, both widely used in cognitive neuroscience, in each of which a decision maker must discriminate between two alternatives, henceforth denoted '1' and '2'. The sensory information itself, as well as its neural representation,

is assumed to be noisy, so that discrimination errors occur. The first task is the *forced-response* paradigm, in which subjects must respond with their best estimate of which alternative (1 or 2) was presented at a fixed time T following stimulus onset. We will also refer to this as the 'interrogation protocol.' Performance on this task is measured by the error rate, or one minus the fraction of correct responses. We note that the interrogation protocol is distinct from the deadlining paradigm (not considered further here), in which the subject is apprised in advance of a fixed, maximal time *before* which all responses must be made.

In the second, *free-response* paradigm, decisions are not demanded at a preset time, but are given when the subject feels that sufficient evidence in favor of one alternative has accumulated. Since the sensory evidence is noisy, response times vary from trial to trial and performance under the free-response condition is characterised by both reaction times and error rates. Here, optimality requires an appropriate balance of speed and accuracy [177, 72, 13].

Following [169] and others, we shall model both these tasks by a pair of competing (mutually inhibitory) neural populations, each of which is selectively responsive to sensory input corresponding to one of the two alternatives. In the forced-response protocol, the neural population with the highest firing rate at time T determines the decision. For free responses, the first of the two populations to cross a firing rate threshold establishes the choice. We do not address the (interesting) question of how thresholds are set or threshold crossings are detected.

6.3.2 Two dimensional nonlinear models and the neural gain parameter

In this section we consider the dynamics of two mutually inhibiting neural populations, each of which receives noisy sensory input from components of the stimulus representing one of the alternatives. We describe two models, both in wide use, for such populations.

The first of these, the leaky integrator *connectionist* model [120, 169], is:

$$\tau_c \frac{dx_1}{dt} = -x_1 - \beta f_{g(t)}(x_2) + a_1(t) + \frac{c(t)}{\sqrt{2}}\eta_t^1, \tag{6.1}$$

$$\tau_c \frac{dx_2}{dt} = -x_2 - \beta f_{g(t)}(x_1) + a_2(t) + \frac{c(t)}{\sqrt{2}}\eta_t^2, \tag{6.2}$$

where the state variables $x_j(t)$ denote the mean input currents to cell bodies of the jth neural population, the integration implicit in the differential equations modelling temporal summation of dendritic synaptic inputs [80]. Additionally, the parameter β sets the strength of mutual inhibition via population firing rates $f_{g(t)}(x_j(t))$, where $f_{g(t)}(\cdot)$ is the sigmoidal 'activation' (or 'frequency-current' or neural 'input-output') function to be described shortly. The stimulus signal received by each population is $a_j(t)$, and the noise terms polluting this signal are $c(t)\eta_t^j$, where $c(t)$ sets r.m.s. noise strength and the η_t^j are (independent) white noise processes with variance $E(\eta_t^j - \eta_{t'}^j)^2 = \delta(t - t')$. The time constant τ_c reflects the rate at which neural activities

decay in the absence of inputs and respond to input changes. Under the free-response paradigm a decision is made and the response initiated when the firing rate $f_{g(t)}(x_j)$ of either population first exceeds a preset threshold θ_j; it is normally assumed that $\theta_1 = \theta_2 = \theta$. For the interrogation protocol, the population with greatest activity (and also firing rate) at time T determines the decision. We also assume that activities decay to zero after response and prior to the next trial, so that the initial conditions for (6.1-6.2) are $x_j(0) = 0$.

The subscript in $f_{g(t)}(\cdot)$ indicates dependence on the time-varying gain, or sensitivity, $g(t)$ of the neural populations: gain sets the slope of the activation function. For example, the logistic function

$$f_{g(t)}(x) = \frac{1}{1 + \exp\left(-4g(t)\left(x - b\right)\right)} = \frac{1}{2}\left[1 + \tanh\left(2g(t)(x - b)\right)\right] \qquad (6.3)$$

has maximal slope $g(t)$ (see Fig. 6.1, left). While this specific form is not required for the results derived below, we do assume that f_g takes its time-dependent maximal slope $g(t)$ at some time-independent point, as for (6.3).

As already mentioned, the connectionist model describes the time evolution of current inputs. We now introduce another model in which the firing rates of neural populations are themselves integrated over time. First we give the linearized version of this *firing rate* model:

$$\tau_c \frac{dy_1}{dt} = -y_1 + f^l_{g(t)}\left(-\beta y_2 + a_1(t) + \frac{c(t)}{\sqrt{2}}\eta^1_t\right), \qquad (6.4)$$

$$\tau_c \frac{dy_2}{dt} = -y_2 + f^l_{g(t)}\left(-\beta y_1 + a_2(t) + \frac{c(t)}{\sqrt{2}}\eta^2_t\right). \qquad (6.5)$$

Here, the y_j are the firing rates of population j and other terms are as above. The linear function

$$f^l_{g(t)}(x) = \frac{1}{2} + g(t)\left(x - b\right), \qquad (6.6)$$

derives from replacing the logistic (or any similar monotonic) function by the linear approximation $f^l_{g(t)}(\cdot)$ around its point of maximal slope. Note that the firing rate y_j of the jth population approaches an equilibrium set by the input currents to this population, passed through the (linearized) frequency-current function. This model must be reformulated to allow for nonlinear functions $f_{g(t)}$, because white noise does not make sense as an argument in such a function, cf. [65]. In particular, we assume that, as in (6.4-6.5), the strength of firing rate fluctuations in response to noise in inputs scales with $g(t)$ (i.e., with the maximal sensitivity of firing rates to the deterministic component of the input). This yields

$$\tau_c \frac{dy_1}{dt} = -y_1 + f_{g(t)}\left(-\beta y_2 + a_1(t)\right) + g(t)\frac{c(t)}{\sqrt{2}}\eta^1_t, \qquad (6.7)$$

$$\tau_c \frac{dy_2}{dt} = -y_2 + f_{g(t)}\left(-\beta y_1 + a_2(t)\right) + g(t)\frac{c(t)}{\sqrt{2}}\eta^2_t, \qquad (6.8)$$
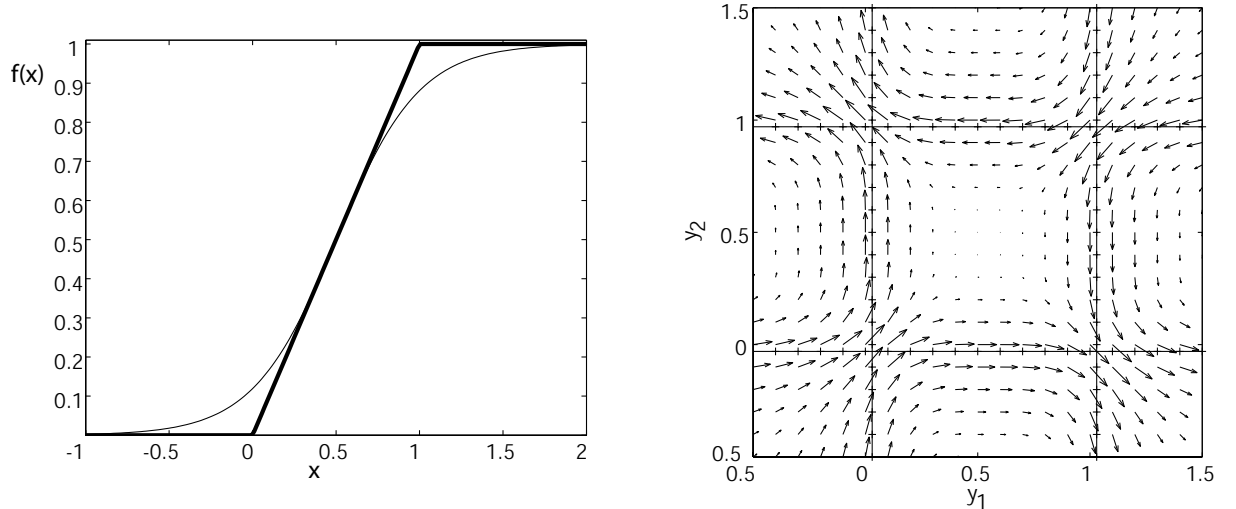
Figure 6.1: (left) Comparison of logistic and piecewise linear activation functions; $g = 1$, $b = 0.5$. (right) Comparison of logistic and piecewise linear vectorfields $F(y_1, y_2)$ and $F^{pw}(y_1, y_2)$ for the piecewise-linear firing rate model (6.10-6.11): the difference $F(y_1, y_2) - F_{pw}(y_1, y_2)$ is plotted. Also shown for reference are the nine phase space tiles described in Figure 6.2. Here additionally $\tau_c = 1$, $\beta = 1$, $a_1 = 1.03$, $a_2 = 0.97$.

which is valid for all $f(\cdot)$ and reduces to the form (6.4-6.5) for linear $f(\cdot)$. Note that the firing rate model (6.7-6.8) is a standard two-unit recurrent neural network with additive noise [87]. As above, we take initial conditions $y_j(0) = 0$, and note that threshold-crossing in the free-response case is detected directly via $y_j = \theta_j$.

For the questions of optimal stimulus processing addressed here, the most important distinction between the connectionist (6.1-6.2) and firing rate (6.4-6.5)-(6.7-6.8) models is whether the inputs $a_j(t) + c(t)/\sqrt{2}\eta_t^j$ enter as separate additive terms, as in the former, or as arguments to the activation function $f_g(t)$, as in the latter. As explained at the end of Section 6.4, this determines whether changes in gain directly adjust the sensitivity of neural units to all inputs or just to feedback from the competing unit, and it results in qualitatively different predictions for optimal gain schedules in the two models. While we expect that future work on low-dimensional descriptions of the population dynamics of spiking neurons (extending, e.g., [26, 184, 132, 157, 49] to include neurotransmitter effects) will result in more refined models, here we study the 'simple' connectionist and firing rate descriptions. Throughout, we use variables x_j in referring to the former and y_j to the latter.

6.3.3 Piecewise linear approximations

As in [169, 21], Eqn. (6.3) may approximated by a piecewise linear function:

$$f_{g(t)}(\xi) \approx f_{g(t)}^{pw}(\xi) = \begin{cases} 0 \text{ for } \xi \in (-\infty, b - \frac{1}{2g}] \\ \frac{1}{2} + g(t)(\xi - b) \text{ for } \xi \in [b - \frac{1}{2g}, b + \frac{1}{2g}] \\ 1 \text{ for } \xi \in [b + \frac{1}{2g}, \infty) \end{cases} , \qquad (6.9)$$

as illustrated in Fig. 6.1.

For ease of reference, we rewrite Eqns. (6.7-6.8) following piecewise linearization:

$$\tau_c \frac{dy_1}{dt} = -y_1 + f_{g(t)}^{pw}(-\beta y_2 + a_1(t)) + g(t)\frac{c(t)}{\sqrt{2}}\eta_t^1 , \qquad (6.10)$$

$$\tau_c \frac{dy_2}{dt} = -y_2 + f_{g(t)}^{pw}(-\beta y_1 + a_2(t)) + g(t)\frac{c(t)}{\sqrt{2}}\eta_t^2 . \qquad (6.11)$$

The difference between the vectorfield of the fully nonlinear model (6.7-6.8) and that of (6.10-6.11) is illustrated in Fig. 6.1 (right) for a specific choice of parameters.

The (y_1, y_2) phase space of the piecewise linear firing rate model (and of the analogous connectionist model) is tiled by nine regions divided by pairs of horizontal and vertical lines at the break points of f_g^{pw}, each having a distinct linear vectorfield: see Fig. 6.2. In the following section, we will describe two cases in which this tiled structure can be used to reduce Eqns. (6.7-6.8) to a one-dimensional system.

6.3.4 Representing decision dynamics in one dimension

As discussed above and in [169], in the forced response protocol, the choice $j = 1$ or 2 is made according to which of the two neural populations has the greatest activity or firing rate at interrogation time T. Therefore, knowledge of the difference

$$x(T) \triangleq x_1(T) - x_2(T) \text{ or } y(T) \triangleq y_1(T) - y_2(T) \qquad (6.12)$$

determines the outcome and reduction of the original two-dimensional problem to a single variable does not *inherently* imply any loss in accuracy. For example, if the difference in firing rates is described by a time-dependent probability density $p(y, t)$ (whose distribution represents variability across behavioral trials), then the error rate at interrogation time T is

$$ER = \int_0^\infty p(y, T)dy \qquad (6.13)$$

if alternative 2 was presented (that is, if $a_2 > a_1$ for $t > t_s$), and

$$ER = \int_{-\infty}^0 p(y, T)dy \qquad (6.14)$$

if alternative 1 was presented. Similar conclusions hold for the connectionist model.

For the free choice protocol the situation is more subtle. The single variable x or y
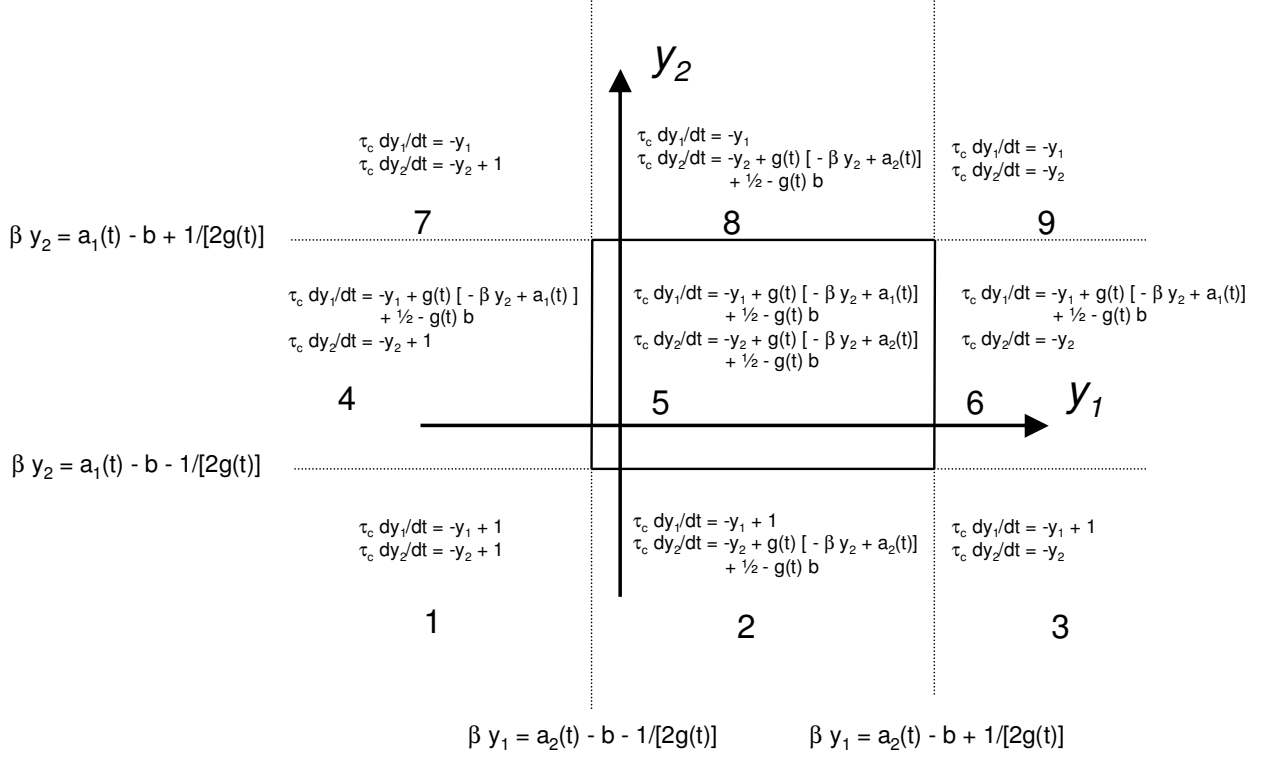
Figure 6.2: The piecewise linear vectorfield of the firing rate model (6.10-6.11). The central tile is surrounded by a solid box.

is sufficient to characterize the decision only if the probability density of solutions to (6.1-6.2) or (6.7-6.8) has approximately collapsed along a one-dimensional 'decision manifold' $\mathcal{M}$ by the time the threshold is crossed; see Fig. 6.3. In this sketch, the decision manifold, parameterized by y, is the unstable, center or weak stable manifold [82] of the indicated fixed point. For collapse to $\mathcal{M}$ to occur, the eigenvalue characterizing dynamics normal to the manifold must be sufficiently negative compared with the other eigenvalue and noise strength c, so that the majority of sample paths cross the thresholds $x_j = \theta$ (or $y_j = \theta$) near their intersections with $\mathcal{M}$ [169, 21, 13]. These requirements are met by two distinct parameter sets to be introduced below.

Dimension reduction and transient gain in two simple cases

In two cases, a simple equation for the evolution of $x(t)$ or $y(t)$ may be derived. These cases are characterized by a dominant proportion of solutions to (6.10-6.11) (i.e., for 'most' realizations of the noise processes $\eta_j(t)$) (i) being confined to a single tile for the duration of the decision process or (ii) 'jumping' together between tiles. The first of these situations occurs for 'case 1' parameter sets, in which, for example, the onset of salience (i.e., $a_1 \neq a_2$) in input currents is accompanied by large transients in the magnitude of these inputs. The second 'case 2' occurs for stimuli in which salience appears without such transients in magnitude. We now consider these cases in detail
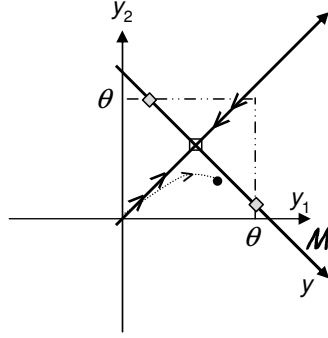
Figure 6.3: Reduction to one dimension. The coordinate y (or x) of Eqn. (6.12) parameterizes the decision manifold $\mathcal{M}$ (see text): the invariant manifold containing the fixed point indicated by the square. In the free response (but not the forced response/interrogation) protocol, collapse of noisy solution trajectories along $\mathcal{M}$ is required for accurate description in one dimension (cf. Figs. 6.4 and 6.5 (right)) so that trajectories (dotted line and point) cross thresholds arbitrarily close to the intersections of $\mathcal{M}$ with the thresholds $y_j = \theta$.

for the firing rate model.

Case 1: Trajectories confined to the central tile, gain parameter directly modulated

The central tile of the firing rate phase plane, where both functions $f_{g(t)}^{pw}(\cdot)$ appearing in Eqns. (6.10-6.11) are linearly increasing, is defined by
$\beta y_1 \in \left[a_2(t) - b - \frac{1}{2g(t)}, \ a_2(t) - b + \frac{1}{2g(t)}\right]$ and $\beta y_2 \in \left[a_1(t) - b - \frac{1}{2g(t)}, a_1(t) - b + \frac{1}{2g(t)}\right]$.
If

$$b - \frac{1}{2g(t)} < a_1(t), \, a_2(t) < b + \frac{1}{2g(t)} \,, \tag{6.15}$$

then the central tile always contains the origin and some part of the first quadrant (note that this quadrant is invariant under the deterministic part of Eqns. (6.7-6.8) if f is non-negative) so that decision dynamics starting at the origin may (for suitable choices of other parameters) take place entirely within the central tile. For example, if $b = 0.5$ and $0 < g(t) \leq 1$, then $a_1(t), a_2(t)$ may take values between 0 and 1 while still satisfying (6.15).

Fig. 6.4 shows a sample of solutions of the piecewise-linearized firing rate model for the piecewise constant parameters $g(t) = \{0.3, \, t < t_s; 1, \, t \geq t_s\}$, $a_1(t) = \{1, t \leq t_s; 1.03, t > t_s\}$, $a_2(t) = \{1, t \leq t_s; 0.97, t > t_s\}$, $c(t) \equiv 0.09\sqrt{2}$, $b = 0.5$, $\tau_c = 1$, $\theta = 0.725$, $t_s = 10$ and $\beta = 1$. Note that stimuli $a_j(t) \neq 0$ are present throughout, but that coherence ($a_1(t) \neq a_2(t)$) appears in the inputs a_j only at $t = t_s$, so that times $t < t_s$ make up the preparatory phase mentioned in the introduction and the situation corresponds to the introduction of coherence into an entirely random pattern. Assuming that decision thresholds are set within the boundaries of the central tile or that the interrogation time T is sufficiently small so that only a negligible pro-
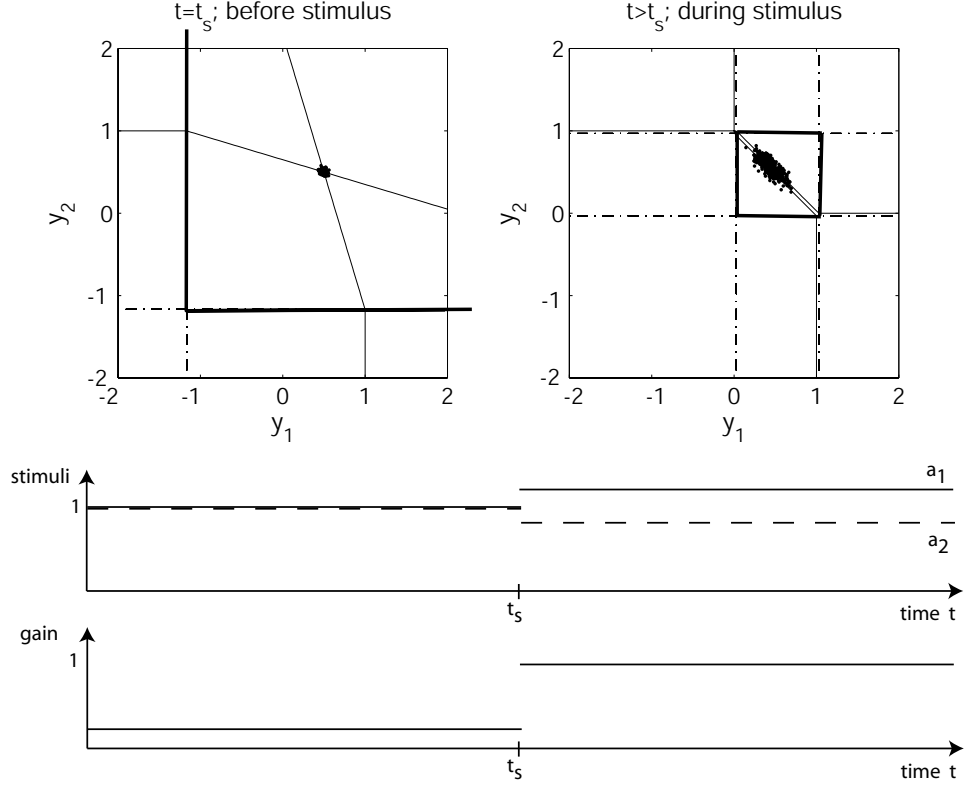
Figure 6.4: Case 1: Solutions confined to central tile. Scatter plot of trajectories both at the end of the preparatory period and hence at the moment of stimulus onset t_s (left) and during the stimulus ($t = t_s + 2$, right). The tiling of the plane is shown with dot-dashed lines; cf. Fig. 6.2; the central tile is outlined in solid and extends outside the plotted domain in the left panel. Parameter values are given in text. Also shown are nullclines for Eqns. (6.10-6.11) as thin solid lines. The lower panels show stimuli $a_j(t)$ and gain $g(t)$ as functions of time.

portion of solutions have left this tile, solutions are effectively confined to the central tile for all times of interest. This behavior characterizes 'case 1' parameter sets, for which subtraction of Eqns. (6.10-6.11) yields the one-dimensional SDE

$$\tau_c \frac{dy}{dt} = -y + g(t)\left(\beta y + a(t)\right) + g(t)c(t)\eta_t \quad \text{(firing rate model)} \ , \qquad (6.16)$$

where we define the net rate of incoming evidence as

$$a(t) = a_1(t) - a_2(t) \,. \qquad (6.17)$$

We note that transient gain values in this case result from modifications to the firing rate function itself, as solutions explore only the central region of this function in which it is practically linear. This is the 'external' mechanism of dynamic gain change discussed in the Introduction.

118

We also note that an analytical expression for the density of reaction times may be derived if the parameters in (6.16) are constant (i.e., $a(t) \equiv a$, $c(t) \equiv c$) and the gain 'balances' the decay: e.g., $g(t) \equiv g = 1$ in (6.16) (see, e.g., [140]). In this case, (6.16) simplifies to a constant drift diffusion process and the probability that a trajectory first escapes the interval $[-\bar{\theta}, \bar{\theta}]$ at a time $RT = \inf\{t : |y(t)| > \bar{\theta}\}$ from initial condition $y(0) = 0$ has density

$$p(RT) = \frac{\pi c^2}{\bar{\theta}^2} e^{-\frac{a^2 RT}{2c^2}} \left(e^{-\frac{\bar{\theta}a}{c^2}} + e^{\frac{\bar{\theta}a}{c^2}} \right) \sum_{k=1}^{\infty} k \sin\left(\frac{k\pi}{2} \right) \exp\left(\frac{-k^2 \pi^2 c^2 RT}{8\bar{\theta}^2} \right) . \quad (6.18)$$

Here $\pm\bar{\theta}$ correspond to the intersections of the decision manifold $\mathcal{M}$ with the thresholds $y_j = \theta$ of the two-dimensional process (Fig. 6.3). Eqn. (6.18) may be extended to account for distributed initial conditions $y(0) \neq 0$ and other generalizations [140], but we do not use such extensions here.

Similar considerations yield the reduction of the connectionist model restricted to its respective central tile:

$$\tau_c \frac{dx}{dt} = -x + \beta g(t)x + a(t) + c(t)\eta_t \quad \text{(connectionist model) .} \quad (6.19)$$

Note that gain multiplies the last three terms in (6.16), but only the second in (6.19).

Case 2: Trajectories switch tiles, changing effective gain

We now consider the case of stimuli $a_j(t)$ that 'suddenly' turn on from zero at time t_s while the gain parameter $g(t) \equiv g$ remains constant, and show how stimulus onset itself can give rise to a time-dependent one-dimensional reduction that resembles the reduction to (6.16) obtained above. This corresponds to appearance of a partially coherent stimulus replacing a fixation spot. Since $a_1(t) = a_2(t) = 0$ for $t \leq t_s$, in this period there is a stable fixed point at $(0,0)$ if $b \geq \frac{1}{2g}$. If $b = \frac{1}{2g}$, the situation simplifies: while $t \leq t_s$, $(0,0)$ lies exactly at the corner of tile 9 (see Fig. 6.2), to which tile solutions are confined (modulo noise effects). At stimulus onset t_s, tile boundaries shift, so that, for appropriate choices of $a_1(t), a_2(t) > \frac{1}{2g(t)} - b$ for $t > t_s$, the origin and the cluster of solutions in its neighborhood at time $t = t_s^+$, suddenly finds itself in the central tile 5. For concreteness, we fix parameters meeting the requirements $b = \frac{1}{2g}$ and $a_1(t) = a_2(t) = 0$ for $t \leq t_s$ as follows: $a_1(t) = \{0, t \leq t_s; 1.03, t > t_s\}$, $a_2(t) = \{0, t \leq t_s; 0.97, t > t_s\}$, $g = 1$ and all other parameters as for the example in case 1. See Fig. 6.5.

To determine the appropriate linear (two- and one-dimensional) reductions for these parameters, we use Eqns. (6.10-6.11) restricted to tile 9 for the preparatory phase $t \leq t_s$, and restricted to tile 5 for times $t > t_s$ during stimulus presentation (we make the same assumptions about the interrogation time or thresholds as for case 1, so that solutions remain in the central tile 5 for all times $t > t_s$ of relevance to the

decision). This yields the one-dimensional equation

$$\tau_c \frac{dy}{dt} = -y + \begin{cases} gc(t)\eta_t & \text{for } t \le t_s \\ g[\beta y + a(t)] + gc(t)\eta_t & \text{for } t > t_s \end{cases} \quad , \tag{6.20}$$

(and an analogous reduction to a linear two-dimensional model).

Equation (6.20) is similar to the reduction (6.16), if the stimulus and gain functions in the latter are piecewise constant, as for the example parameters of case 1. The major difference is that the noise coefficient remains constant for (6.20). As we see in the next section, the statistics produced by the one-dimensional models (6.16) and (6.20) can nevertheless agree rather well. Thus, transient gain strategies to be derived for the more general (6.16) in Section 6.4 can be approximately implemented for stimuli undergoing large steps, with no changes in the gain of the activation functions per se.

Similar considerations hold for 'case 1' and 'case 2' reductions of the connectionist model, but we do not pursue this here.

Finally, we note that [20] contains an explicit comparison of error rates and reaction times predicted by the two-dimensional firing rate model with logistic activation functions $f_{g(t)}$, the two-dimensional model with piecewise-linear activation functions $f_{g(t)}^{pw}$ (6.10-6.11), and the one-dimensional reduction (6.16). This comparison verifies the expected agreement between one- and two-dimensional models for the parameters chosen in Case 1 and Case 2 above, adding to the evidence in [21, 169, 13] that one-dimensional linearized models can capture the essential decision making dynamics for a reasonably broad range of model parameters.

6.3.5   Drift-diffusion and the one dimensional models as linear filters

We introduce a third one-dimensional SDE, an extension of the drift-diffusion model of [113, 139] in which both drift and diffusion terms are multiplied by a common gain factor $g(t)$:

$$\tau_c \frac{dz}{dt} = g(t)[a(t) + c(t)\eta_t] \quad \text{((pure) drift-diffusion model)} \ . \tag{6.21}$$

Eqn. (6.21) and the one-dimensional reductions of the firing rate and connectionist equations (6.16) and (6.19) are Ornstein-Uhlenbeck processes, (affine-) linear in the activities x, y, and z and in the input

$$\underbrace{I(t)}_{input} = \underbrace{a(t)}_{signal} + \underbrace{c(t)\eta_t}_{noise} \ . \tag{6.22}$$
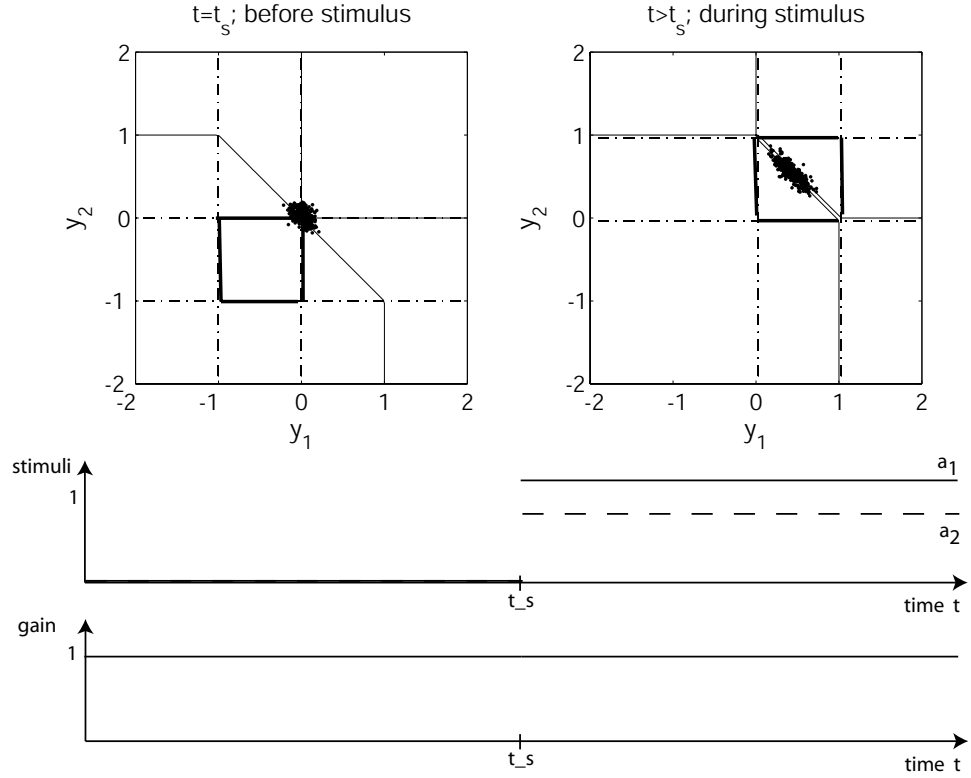
Figure 6.5: Case 2: Trajectories switch tiles. Scatter plot of trajectories both at the end of the preparatory period and hence at the moment of stimulus onset t_s (left) and during the stimulus ($t = t_s + 2$, right). The tiling of the plane is shown with dot-dashed lines; cf. Fig. 6.2; the central tile is outlined in solid. Parameter values are given in text. Also shown are nullclines for Eqns. (6.10-6.11) as thin solid lines. The lower panels show stimuli $a_j(t)$ and gain $g(t)$ as functions of time.

We may explicitly solve all these SDEs, for a given realization of the white noise process η_s, $s \in [0, t]$, to obtain respectively

$$z(t) = \int_0^t \frac{g(s)a(s)}{\tau_c} ds + \int_0^t \frac{g(s)c(s)}{\tau_c} dW_s \tag{6.23}$$

for the drift diffusion model,

$$x(t) = \int_0^t \frac{a(s)}{\tau_c} \exp\left(\frac{1}{\tau_c} \int_s^t [\beta g(s') - 1] \, ds'\right) ds + \int_0^t \frac{c(s)}{\tau_c} \exp\left(\frac{1}{\tau_c} \int_s^t [\beta g(s') - 1] \, ds'\right) dW_s \tag{6.24}$$

for the connectionist model, and

$$y(t) = \int_0^t \frac{a(s)g(s)}{\tau_c} \exp\left(\frac{1}{\tau_c} \int_s^t [\beta g(s') - 1] \, ds'\right) ds + \int_0^t \frac{c(s)g(s)}{\tau_c} \exp\left(\frac{1}{\tau_c} \int_s^t [\beta g(s') - 1] \, ds'\right) dW_s \tag{6.25}$$

for the firing rate model. Here, dW_s is an increment of a Wiener process, of which the white noise process η_s is the formal time derivative, and we have assumed unbiased initial data $x(0) = y(0) = z(0) = 0$. These expressions all take the form

$$w(t) = \int_0^t K(t,s)a(s)ds + \int_0^t K(t,s)c(s)dW_s \,, \tag{6.26}$$

and so we conclude that (6.23-6.25) all compute linear filters of their inputs.

At any fixed time t, $w(t)$ is a gaussian-distributed random variable with mean $\int_0^t K(t,s)a(s)$ and variance $\int_0^t K^2(t,s)c^2(s)ds$. Using this fact, after a change of variables the error rate expression (6.14) becomes

$$\mathrm{ER} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\left|\int_0^t K(t,s)a(s)\right|}{\int_0^t K^2(t,s)c^2(s)ds}\right)\right] \,. \tag{6.27}$$

6.4 Optimal signal discrimination in the one- dimensional models

We now ask what functional form of $g(t)$ optimizes performance for Eqns (6.23-6.25), thereby computing optimal gain trajectories for the (reduced) drift-diffusion, connectionist, and firing rate models.

6.4.1 Optimal statistical tests

Given only the noisy input function (6.22), consider the task of deciding whether $I(t)$ was generated by time-dependent signals $a_0(t)$ or $a_1(t)$: hypotheses 0 and 1, resp. This can be accomplished in two distinct ways, mirroring the interrogation and free response protocols of Section 6.3. In the first, the decision is made at a fixed time T; in the second, it is made when some preset level of confidence is reached. Optimal performance in the first version of the task implies that as few errors as possible are made; in the second it implies that the decision must be made as quickly as possible for a fixed error tolerance, timed from stimulus onset at time $t = 0$. The best strategy in the first version is the (continuum limit of the) Neyman-Pearson test; in the second version it is the sequential probability ratio test (SPRT) [172, 114]. Both tests compute an evolving estimate of the log likelihood ratio:

$$l(t) = \log\left[\frac{p\left(\{I(s)|a_0(s), s \in [0,t]\}\right)}{p\left(\{I(s)|a_1(s), s \in [0,t]\}\right)}\right] \stackrel{\triangle}{=} \log\left[\frac{p_0\left(\{I(s), s \in [0,t]\}\right)}{p_1\left(\{I(s), s \in [0,t]\}\right)}\right] \,. \tag{6.28}$$

(the base of the logarithm is arbitrary). In the Neyman-Pearson test, hypothesis 0 is chosen if $l(T) > 0$ and hypothesis 1 if $l(T) < 0$; in the SPRT, hypothesis 0 (resp. 1) is chosen when $l(t)$ first crosses threshold θ (resp. $-\theta$), θ being determined by the error tolerance.

Writing the input $I(t)$ (6.22) as a sum of its increments for an appropriate discretization of time $\{t^j\}$:

$$I(t) = \sum_j dI^j = \sum_j a(t^j)dt + c(t^j)dW_t^j , \qquad (6.29)$$

we obtain

$$l(t) = \sum_j \log \left[\frac{p_0(dI^j)}{p_1(dI^j)} \right] . \qquad (6.30)$$

Now restrict to the special case in which $a_0(t) = -a_1(t) = a(t)$ and consider the likelihood distributions (now themselves time-dependent) that correspond to an increment $dI(t) = a(t)dt + c(t)dW_t$. Since the dW_t are normally distributed with mean 0, variance dt, we have

$$p_0(t)(dI(t)) = \frac{1}{\sqrt{2\pi c^2(t)dt}} e^{-(dI(t)+a(t)dt))^2/(2c^2(t)dt)} , \qquad (6.31)$$

$$p_1(t)(dI(t)) = \frac{1}{\sqrt{2\pi c^2(t)dt}} e^{-(dI(t)-a(t)dt))^2/(2c^2(t)dt)} . \qquad (6.32)$$

The corresponding increment of likelihood evidence to (6.28) is

$$dl_t = \log \left(\frac{p_1(dI_t)}{p_0(dI_t)} \right) = k\frac{a(t)}{c^2(t)}dI_t , \qquad (6.33)$$

where $k = 2\log(e)$ depends on the base of the logarithm. Substituting for dI_t, we obtain a differential equation for the total evidence l_t accumulated at time t,

$$dl_t = k \left[\frac{a^2(t)}{c^2(t)} dt + \frac{a(t)}{c(t)} dW_t \right] , \qquad (6.34)$$

which may be integrated to yield:

$$l(t) = \int_0^t k\frac{a^2(s)}{c^2(s)}ds + \int_0^t k\frac{a(s)}{c(s)}dW_s . \qquad (6.35)$$

Comparing with Eqn. (6.26) shows that the optimal filter is

$$K(t,s) = k\frac{a(s)}{c^2(s)} : \qquad (6.36)$$

this is the matched filter for white noise which is fundamental in signal processing [136]. Note that, in (6.34-6.35) only the signal-to-noise ratio (a/c) appears.

6.4.2 A direct proof that the kernel $K(t,s) = k\frac{a(s)}{c^2(s)}$ is optimal in the interrogation paradigm

As follows from its matched filter property, the linear filter $K(t,s) = k\frac{a(s)}{c^2(s)}$ which computes log likelihood $l(t)$ for inputs with white noise also produces, for all times t, a filtered (and gaussian) version $w(t)$ of the input (Eqn. (6.26)) with a maximal integrated signal-to-noise ratio

$$F[K;a,c](t) = \frac{\left| \int_0^t K(t,s)a(s)ds \right|}{\sqrt{\mathbb{E}\left(\int_0^t K(t,s)c(s)dW_s \right)^2}} = \frac{\left| \int_0^t K(t,s)a(s)ds \right|}{\sqrt{\int_0^t K^2(t,s)c^2(s)ds}}. \tag{6.37}$$

For completeness, we now demonstrate this directly.

Minimization of the error rate (6.13) or (6.14) for (fixed) interrogation at time $t = T$ is achieved by maximizing F over all possible kernels $K(s)$. This problem in the calculus of variations is solved by computing the first and second variations, with respect to K, of the functional F, setting the first to zero to determine a candidate $\bar{K}$ for the optimal K, and evaluating the second at $\bar{K}$ to check that $D_K^2 F$ is negative (semi-) definite. Henceforth we drop explicit reference to the (fixed, arbitrary) interrogation time $t = T$ in the function K and write $K(T,s) = K(s)$. We compute:

$$\frac{\delta F}{\delta K} = \lim_{\epsilon \to 0} \frac{d}{d\epsilon} F[K + \epsilon\gamma; a, c](T)$$

$$= \lim_{\epsilon \to 0} \frac{d}{d\epsilon} \left\{ \frac{\int_0^T a(s)[K(s) + \epsilon\gamma(s)]\,ds}{\left[2\int_0^T c^2(s)[K^2(s) + 2\epsilon g(s)\gamma(s) + \epsilon^2\gamma^2(s)]\,ds \right]^{\frac{1}{2}}} \right\}$$

$$= \lim_{\epsilon \to 0} \frac{1}{\sqrt{2}} \left\{ \frac{\int_0^T a(s)\gamma(s)\,ds}{[H(T,\epsilon)]^{\frac{1}{2}}} - \frac{\int_0^T a(s)[K(s) + \epsilon\gamma(s)]\,ds \int_0^T c^2(s)[K(s)\gamma(s) + \epsilon\gamma^2(s)]\,ds}{[H(T,\epsilon)]^{\frac{3}{2}}} \right\}$$

$$= \frac{\int_0^T a(s)\gamma(s)\,ds \int_0^T c^2(s)K^2(s)\,ds - \int_0^T a(s)K(s)\,ds \int_0^T c^2(s)K(s)\gamma(s)\,ds}{\sqrt{2}\left[\int_0^T c^2(s)K^2(s)\,ds \right]^{\frac{3}{2}}}, \tag{6.38}$$

where $H(T,\epsilon) = \int_0^T c^2(s)[K^2(s) + 2\epsilon K(s)\gamma(s) + \epsilon^2\gamma^2(s)]\,ds$. Setting (6.38) equal to zero and using the fact that the variation $\gamma(s)$ is arbitrary, we conclude that the critical point indeed occurs at $\bar{K}(s) = k\frac{a(s)}{c^2(s)}$, as given by (6.36).

To compute the second derivative we differentiate the expression within braces in the penultimate step of (6.38) with respect to ϵ once more, set $\epsilon = 0$, and evaluate

the resulting expression at the critical point (6.36), obtaining:

$$\left.\frac{\delta^2 F}{\delta K^2}\right|_{K=\bar{K}} = -\frac{\int_0^T c^2(s)\bar{K}^2(s)\,ds \int_0^T c^2(s)\gamma^2(s)\,ds - \left(\int_0^T c^2(s)\bar{K}(s)\gamma(s)\,ds\right)^2}{\sqrt{2}\left[\int_0^T c^2(s)\bar{K}^2(s)\,ds\right]^{\frac{3}{2}}} \leq 0\,.$$

(6.39)

In the last step we appeal to Schwarz's inequality. This proves that the second variation is negative semidefinite, and vanishes identically only for variations $\gamma(s) = \kappa\bar{K}(s)$ in the direction of $\bar{K}$ (as expected from (6.36), which contains the arbitrary 'scaling' parameter k).

Substituting (6.36) into (6.37) we obtain

$$F[\bar{g}; a, c](T) = \sqrt{\frac{1}{2}\int_0^T \frac{a^2(s)}{c^2(s)}\,ds}\,,$$

(6.40)

and using (6.27), we obtain the minimum possible error rate for interrogation at time t:

$$\text{ER} = \frac{1}{2}\left[1 - \text{erf}\left(\sqrt{\frac{1}{2}\int_0^T \frac{a^2(s)}{c^2(s)}\,ds}\right)\right]\,.$$

(6.41)

Since the integrand $(a/c)^2$ is non-negative, the error rate continues to decrease or at worst remains constant as T increases.

6.4.3   Optimal gains for the three models

We may now extract explicit expressions for optimal gains by setting $K(s) = \bar{K}(s)$ in (6.26) and comparing the resulting integrands with those in the SDE solutions (6.23-6.25).

Pure drift-diffusion model

Comparing (6.26) with (6.23), we see that the optimal gain is simply $\bar{K}$:

$$\bar{g}_{dd}(s) = \tau_c\bar{K}(s) = \tau_c k\frac{a(s)}{c^2(s)}\,;$$

(6.42)

thus, there is a continuum of optimal schedules differing only by a multiplicative scale factor.

Connectionist model

Equations (6.26) and (6.24) give

$$\tau_c\bar{K}(s) = \tau_c k\frac{a(s)}{c^2(s)} = \exp\left(\frac{1}{\tau_c}\int_s^T [\beta\bar{g}_c(s') - 1]\,ds'\right)\,,$$

(6.43)

where $\bar{g}_c$ is the optimal gain for the connectionist model. Taking the log of this expression, differentiating with respect to s, and solving for $\bar{g}_c(s)$, we obtain:

$$\bar{g}_c(s) = \frac{1}{\beta}\left[1 - \tau_c \frac{d}{ds}\log\left(\frac{a(s)}{c^2(s)}\right)\right] . \qquad (6.44)$$

Note that $\bar{g}_c$ is unique and in particular, independent of k and of the interrogation time T. However, $\bar{g}_c$ is not required to be positive, so may not always be physically admissable. The form of $\bar{g}_c$ may be interpreted as follows. When $\left(\frac{a(s)}{c^2(s)}\right)$ is decreasing, $\bar{g}_c(s) > 1/\beta$ and the O-U process (6.19) is unstable; hence solutions 'run away,' in the direction $x(s)$, emphasizing higher-fidelity information that was previously collected. When $\left(\frac{a(s)}{c^2(s)}\right)$ is increasing, $\bar{g}_c(s) < 1/\beta$, the O-U process is stable, and the linear term in (6.19) is attractive, thereby discounting previously integrated information in favor of the higher-fidelity input currently arriving.

We note that, because the 'output' neural activity is determined by a gain-dependent function of the dynamical variable x in the connectionist model (see text following Eqns. (6.1-6.2)), transient gain schedules also adjust the position of free-response thresholds with respect to x. We leave an exploration of this effect, which does not enter the interrogation protocol or affect the firing rate model, for future studies.

Firing rate model

Equations (6.26) and (6.25) give

$$\tau_c \bar{K}(s) = \tau_c k \frac{a(s)}{c^2(s)} = \bar{g}_f(s)\exp\left(\frac{1}{\tau_c}\int_s^T [\beta\bar{g}_f(s') - 1]\,ds'\right) . \qquad (6.45)$$

Defining $f(s) = \tau_c k \frac{a(s)}{c^2(s)} e^{\frac{1}{\tau_c}(T-s)}$, differentiating with respect to s, and restricting to positive functions $\bar{g}_f$, a and c^2 (which we justify below), (6.45) yields

$$\begin{aligned}
f'(s) &= \frac{d}{ds}\left[\bar{g}_f(s)\exp\left(\frac{1}{\tau_c}\int_s^T \beta\bar{g}_f(s')ds'\right)\right] \\
&= \bar{g}_f'(s)\exp\left(\frac{1}{\tau_c}\int_s^T \beta\bar{g}_f(s')ds'\right) - \frac{\beta}{\tau_c}\bar{g}_f^2(s)\exp\left(\frac{1}{\tau_c}\int_s^T \beta\bar{g}_f(s')ds'\right) \\
&= \bar{g}_f'(s)\frac{f(s)}{\bar{g}_f(s)} - \frac{\beta}{\tau_c}\bar{g}_f(s)f(s) . \qquad (6.46)
\end{aligned}$$

Rewriting (6.46), we obtain

$$\begin{aligned}
\frac{d\bar{g}_f(s)}{ds} &= \frac{\beta}{\tau_c}\bar{g}_f^2(s) + \bar{g}_f(s)\frac{f'(s)}{f(s)} = \frac{\beta}{\tau_c}\bar{g}_f^2(s) + \bar{g}_f(s)\frac{d}{ds}\log\left(f(s)\right) \\
&= \frac{\beta}{\tau_c}\bar{g}_f^2(s) + \bar{g}_f(s)\left[\frac{d}{ds}\log\left(\frac{a(s)}{c^2(s)}\right) - \frac{1}{\tau_c}\right] . \qquad (6.47)
\end{aligned}$$

Thus, the condition for optimal gain in the linearized firing rate model is a differential equation, unlike the algebraic relationships for the drift-diffusion and connectionist cases. Note that solutions to (6.47) initialized at positive values remain positive for all time, since the equation has an equilibrium at $\bar{g}_f = 0$, preventing passage through this point. This justifies our assumption of positive $\bar{g}_f$ above and ensures that the optimum gain is 'physical' this sense. In fact, (6.47) may be solved explicitly using the integrating factor $I(s) = \exp\left(\int_0^s l(s')ds'\right)$, where $l(s') \triangleq \frac{d}{ds'}\log\left(\frac{a(s')}{c^2(s')}\right) - \frac{1}{\tau_c}$, yielding

$$\bar{g}_f(s) = \frac{\exp\left(\int_0^s l(s')ds'\right)}{\frac{\beta}{\tau_c}\int_0^s\left[\exp\left(\int_0^{s'} l(s'')ds''\right)\right]ds' + \frac{1}{g(0)}}. \tag{6.48}$$

The integral equation (6.45) specifies only an *arbitrary*, positive final condition $\bar{g}_f(T) = k\frac{a(T)}{c^2(T)}$ for (6.47), since k is itself arbitrary. Any solution of (6.47) with positive initial condition (as long as it is defined) therefore delivers a member of the continuum of optimal gain functions for the linearized firing rate model. This is in striking contrast to the unique optimal gain (6.44) in the connectionist model, and, since the different $\bar{g}_f$ generally have different forms (see below), it also contrasts with the multiplicity of 'scaled' optimal drift-diffusion gain functions (6.42).

Numerical examples

Example 1: We first take constant signal $a(s) \equiv a = 0.06$ and constant noise strength $c(s) \equiv 0.09$ with $\tau_c = \beta = 1$. Then, Eqn. (6.42) gives the family of optimal constant gain functions for the pure drift-diffusion model,

$$\bar{g}_{dd}(s) \equiv \tau_c k a, \tag{6.49}$$

and Eqn. (6.44) gives the unique optimal gain for the connectionist model, again a constant:

$$\bar{g}_c(s) \equiv \frac{1}{\beta}. \tag{6.50}$$

For the same parameter values, the firing rate model gain ODE (6.47) becomes

$$\frac{d}{ds}\bar{g}_f(s) = \frac{\beta}{\tau_c}\bar{g}_f^2(s) - \frac{1}{\tau_c}\bar{g}_f(s). \tag{6.51}$$

Initial conditions $\bar{g}_f(0) \in [0, 1/\beta]$ decay to the fixed point at $\bar{g}_f = 0$, while for $\bar{g}_f(0) > 1/\beta$, gain functions increase to ∞ in finite time. The initial condition $\bar{g}_f(0) = 1/\beta$ yields the constant gain function $\bar{g}_f(s) \equiv 1/\beta$, for which the linearized firing rate model again becomes constant drift Brownian motion: see Fig. 6.6. As expected, all gain profiles produced optimal performance (with 82.7% correct responses returned at interrogation time $T = 2$).

Example 2: We now assume that signal amplitude is zero up to stimulus presentation at time t_s and rises exponentially toward $\bar{a}$ thereafter: $a(s) = \bar{a}[1 - e^{-r(s-t_s)}]$ for $s > t_s$.

Figure 6.6: Optimal gains for constant signal strength $a(s) \equiv 0.09$ (solid line in bottom panel) and constant noise amplitude $c(s) \equiv 0.09$ (dotted line). Top panel: three optimal gain schedules $\bar{g}_f$ solving (6.47); note that these include, but are not limited to, $\bar{g}_f(s) \equiv 1/\beta$ (here $\beta = 1$). Central panel: the unique optimal gain function $\bar{g}_c(s) \equiv 1/\beta$ for the connectionist model, given by Eqn. (6.44).

This form is motivated by the saturating dynamics of input layers which feed forward to decision units in simple connectionist models. We set $\bar{a} = 0.06$, $r = 10$, $t_s = 1$ and take constant noise strength $c(s) \equiv 0.09$ and $\tau_c = \beta = 1$ as previously: see Fig. 6.7 (bottom). As $r \to \infty$, $a(s)$ approaches the piecewise constant functions of Section 6.3.4, for which the one-dimensional reduction is shown to be an adequate model in [20].

For the pure drift-diffusion model, Eqn. (6.42) gives

$$\bar{g}_{dd}(s) = \tau_c k a(s)\,, \tag{6.52}$$

so that, as above, optimal gain trajectories are scaled versions of the signal strength and, in particular, $\bar{g}(s) = 0$ for $s \le t_s$. For the connectionist and firing rate models, however, the formulae (6.44) and (6.47) are valid only while $a(s) > 0$, and additional reasoning is needed to determine optimal gain values in the pre-stimulus period $s < t_s$. For the connectionist model, the integral equation (6.43) is clearly satisfied for $a(s) = 0$ if $g_c(s) = -\infty$, so we set $\bar{g}_c(s) = -\infty$, $s \le t_s$. Since for a 'physical' neural network, activation functions $f_{g(t)}(\cdot)$ are nondecreasing, such negative gain values are

128

Figure 6.7: Optimal gains for exponentially asymptoting signal strength $a(s)$ (solid line in bottom panel) and constant noise amplitude $c(s) \equiv 0.09$ (dotted line). Top panel: three optimal gain schedules $\bar{g}_f$ for the firing rate model solving (6.47) (solid curves); the non-optimal constant gain $g \equiv 1/\beta$ is shown as dot-dashed for reference. The lowest of the solid $\bar{g}_f$'s displays the rise-decay form discussed in the text. Central panel: the unique optimal gain function for the connectionist model, given by Eqn. (6.44); $\bar{g}_c(s) = -\infty$ for $s \leq t_s$.

not directly relevant to biological applications, but illustrate the demand that relative activation x be clamped at zero before the stimulus arrives. As before, we define $\bar{g}_c(s)$ via (6.44) for $s > t_s$. That is, for $t > t_s$,

$$\bar{g}_c(s) = \frac{1}{\beta} \left[1 - \tau_c l(s) \right] , \tag{6.53}$$

where $l(s) = \frac{d}{ds} \log \left(\frac{a(s)}{c^2(s)} \right) = \frac{r}{e^{r(s-t_s)}-1}$ decays from ∞ to 0 as time s increases.

For the firing rate model, we also appeal directly to the integral equation (6.45) to define $g_f(s)$ when $a(s) = 0$. Since (6.45) is satisfied by $\bar{g}_f(s) = 0$, we assume this for $s \leq t_s$. We then determine $\bar{g}_f(s)$ for $s > t_s$ from (6.47), allowing a discontinuity at t_s and taking arbitrary 'initial' conditions $\bar{g}_f(t_s)$. Fig. 6.7 illustrates several optimal functions arising from different choices of $\bar{g}_f(t_s)$. The following fact is helpful in understanding positive solutions of (6.47): orbits lying below $\frac{1}{\beta}[1 - \tau_c l(s)]$ at any
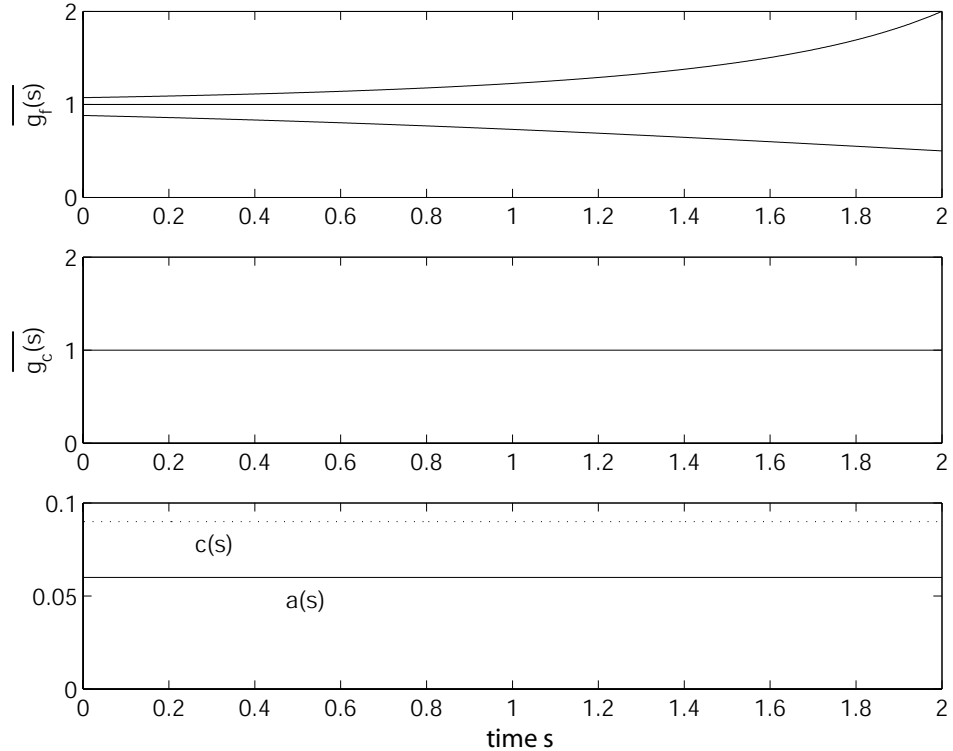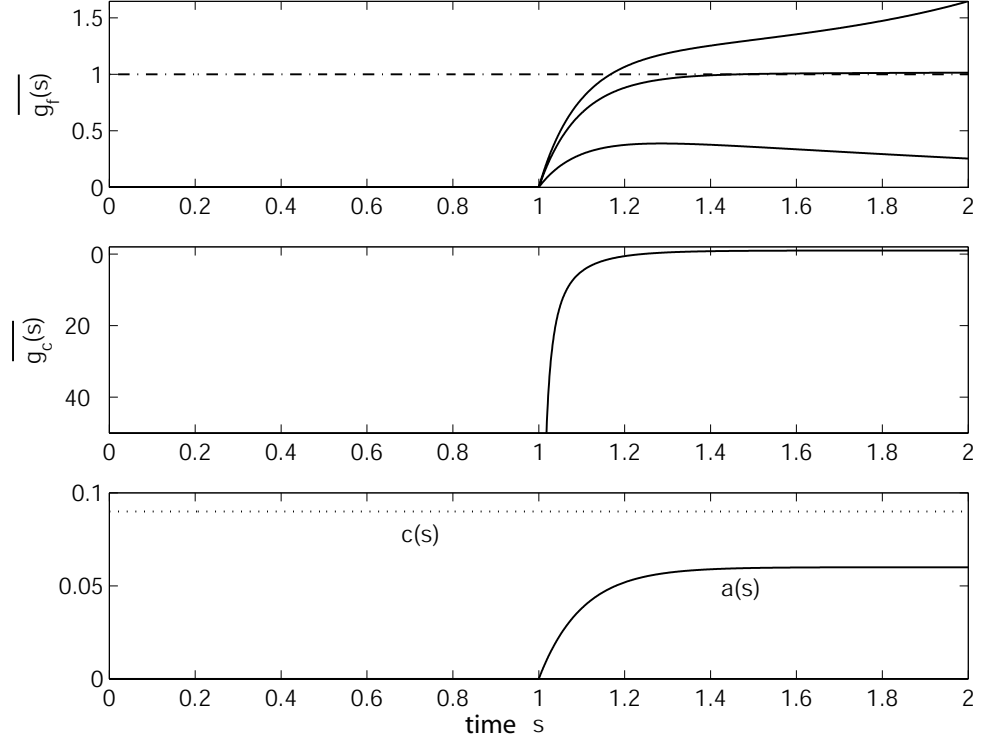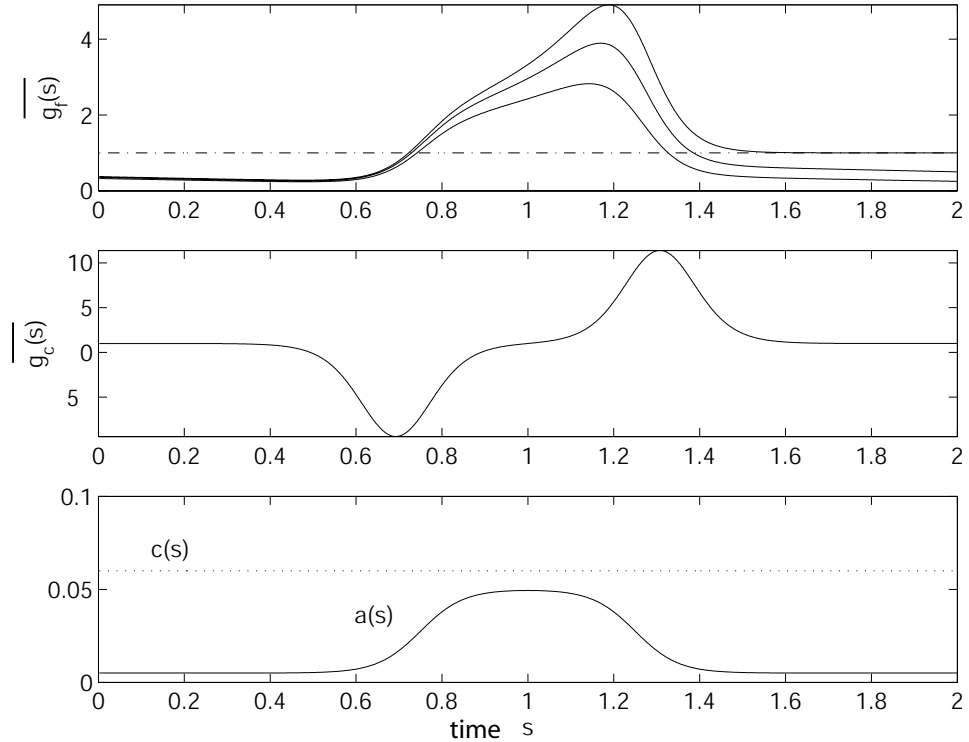
Figure 6.8: Optimal gains for pulsed signal strength $a(s)$ (solid line in bottom panel) and constant noise amplitude $c(s) \equiv 0.06$ (dotted line). Top panel: three optimal gain schedules $\bar{g}_f$ for the firing rate model solving (6.47) (solid curves); the non-optimal constant gain function $g \equiv 1/\beta$ is shown dot-dashed for reference. Central panel: the unique optimal gain function for the connectionist model, given by Eqn. (6.44).

time s decrease toward 0; those above this value increase. Since $\frac{1}{\beta}[1 - \tau_c\, l(s)] \to \frac{1}{\beta}$ as $s \to \infty$, $\frac{1}{\beta}$ asymptotically forms a separatrix between optimal gain trajectories that decay and those that diverge to ∞. Also, note that the Case 2 parameters for the two-dimensional firing rate model of Section 6.3.4 implement a step in effective gain values up to $1/\beta = 1$, so that in this case nearly optimal signal processing occurs with no explicit adjustment of the gain parameter. The performance resulting from optimal gain trajectories in all models is 73.1% correct responses at interrogation at time $T = 2$; for comparison, the (non-optimal) constant gain $\bar{g}_f(s) \equiv 1/\beta$ produces only 66.4% correct.

Gains must remain bounded for all time to be of practical interest. A family of optimal gain schedules of this form, determined by their (sufficiently small) initial conditions, will always exist for monotonically rising and bounded stimuli $a(s)$ such as that chosen here. As we elaborate in Section 6.5, their 'rise-decay' pattern resembles the gain produced by dissipating pulses of the neuromodulator norepinephrine delivered to cortical decision areas via the locus coeruleus, hence providing a clue that this brainstem organ may be assisting near-optimal decision making.

130

Example 3: We finally assume that $a(s)$ smoothly increases from a low to a higher level and then returns to its original level, corresponding to a transient increase in stimulus salience. We model this as a difference of two sigmoids: $a(s) = a_0 + \frac{\bar{a}}{1+\exp(-4r(t_{s,1}-s))} - \frac{\bar{a}}{1+\exp(4r(t_{s,2}-s))}$, with parameters $a_0 = -.04$, $\bar{a} = .045$, $t_{s,1} = 0.75$, $t_{s,2} = 1.25$, and $r = 20$: see Figure 6.8. Additionally, we take constant noise strength $c(s) \equiv 0.06$ and $\tau_c = \beta = 1$.

For the pure drift-diffusion model, Eqn. (6.42) again gives $\bar{g}_{dd}(s) = \tau_c k a(s)$, and for the connectionist and firing rate models, we may use (6.44) and (6.47) for the entire time interval of interest since $a(s)$ is strictly positive. The resulting optimal gain trajectories, shown in Fig. 6.8, yield 70.8% correct responses at interrogation time $T = 2$, compared with 64.9% correct obtained for constant gain $g_f(s) \equiv 1/\beta$ in the firing rate model. Note that the form of the optimal $\bar{g}_c(s)$ illustrates the intuitive explanation given in Section 6.4.3: when the signal-to-noise ratio increases, $\bar{g}_c(s)$ *decreases*, suppressing previously integrated information, and vice-versa.

In summary, we have shown in this section that the gain schedules yielding optimal performance in (reduced) neural models of decision tasks depend strongly on the time course of task stimuli as well as the structure (i.e., firing rate vs. connectionist) of the underlying model. These optimal gain schedules implement matched filters, maximizing the signal-to-noise ratio in the difference between activities of neural populations representing competing task alternatives. For systems well described by the connectionist model, neural mechanisms may be expected to depress the gain (i.e., strength of inhibitory feedback) below the 'balanced' level of $1/\beta$ when stimulus salience is increasing and enhance gain above this level when salience is decreasing. However, for the firing rate model an optimal decision network may 'choose' among a variety of gain schedules of qualitatively different forms. One neurobiological implication of this flexibility is explored in the following section.

6.5 The locus coeruleus brainstem area and optimal gain trajectories

Neurons comprising the brainstem nucleus locus coeruleus (LC) emit the neurotransmitter norepinephrine (NE) to targets widely distributed throughout the brain, including cortical areas involved in decision tasks. While NE has disparate and complex effects on different brain regions, a dominant cortical role is believed to be modulation of neuronal gain at both the single cell and population levels [167, 154]. Recordings of cortical neuron responses to stereotyped inputs at various latencies following activation of LC reveal these gain effects: responses to a fixed input are larger (in certain experimental ranges) following LC activation than in control recordings without LC, and this elevated sensitivity decays with a time constant $\tau_{NE} \approx 0.2$ sec [175].

Since the firing rate of LC neurons governs NE release rate, we propose the fol-
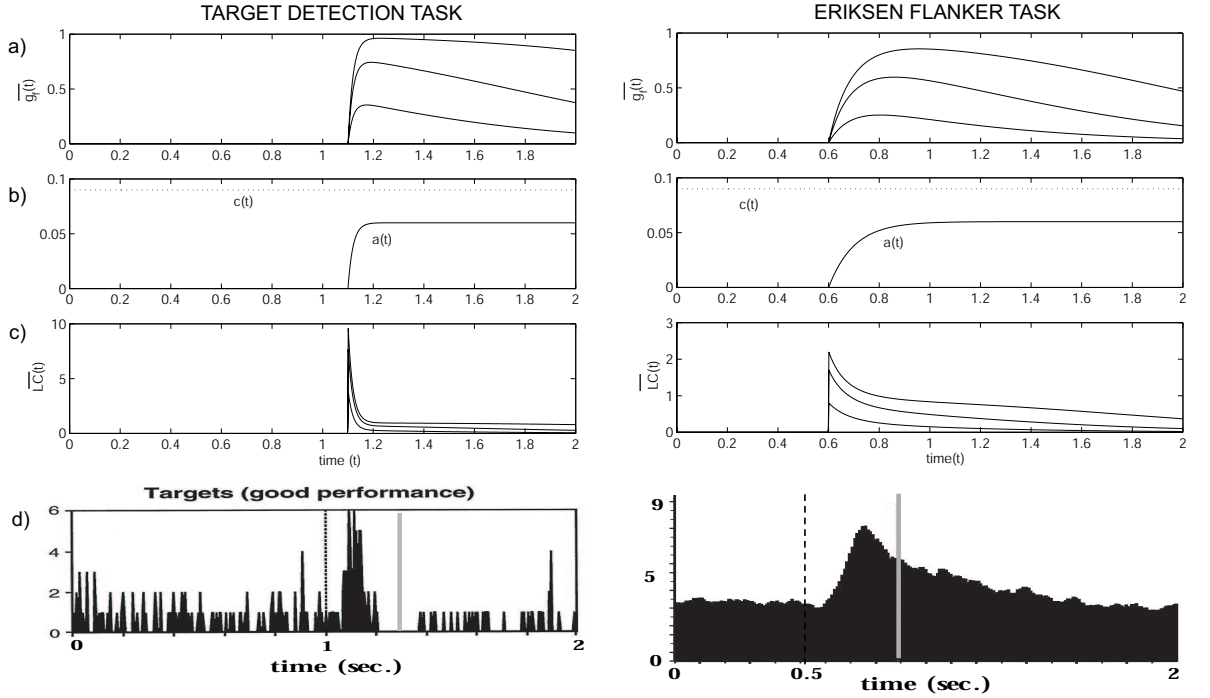
Figure 6.9: Comparison of optimal gain theory with empirical data for two psychological tasks. (a) Optimal gain schedules for the firing rate model, for rapid (left) and gradual (right) onset of stimulus $a(t)$ to neural units (with a processing time lag of 0.1 sec following sensory cue), as shown in (b). (c) The corresponding optimal time courses of LC firing rate. (d) Histograms of LC firing rates recorded in the two tasks: (left), the target detection task [167] and (right), the Eriksen flanker task, with data kindly provided by the authors of [34]. Vertical dashed lines indicate onset of sensory stimuli, and vertical grey (solid) lines indicate mean behavioral reaction time (standard deviations are $\approx$ 34 and 114 msec. for the target detection and Eriksen tasks, respectively).

lowing simple model for cortical gain $g(t)$:

$$\tau_{NE}\,\dot{g}(t) = k_{LC}\,LC(t) - g(t)\,. \tag{6.54}$$

Here, $LC(t)$ denotes the time-dependent rate of LC firing and k_{LC} is a constant relating this rate to equilibrium values of cortical gain. This model's limitations in describing the underlying biology include the fact that $g(t)$ decays to zero in the absence of LC firing (this could be rectified by adding a constant 'gain floor' g_{base}). Nevertheless, it allows us to make an interesting qualitative point in relating recent data on LC firing rates to optimal strategies for the processing of noisy sensory stimuli. Inverting (6.54) and inserting an optimal gain trajectory yields a prediction for the

optimal time course of LC activity:

$$\overline{LC}(t) = \frac{1}{k_{LC}} \left(\tau_{NE}\, \dot{\bar{g}}(t) + \bar{g}(t) \right) . \tag{6.55}$$

Fig. 6.9(d) shows histograms of LC firing rates recorded from monkeys performing two different psychological tasks: target identification, in which a horizontal or vertical bar must be detected, and the Eriksen flanker task, in which a central cue must be identified while an array of distractors is ignored. Since the second task involves more complex stimulus processing, we assume as in [24] that the onset of stimulus representation in cortical decision areas is more gradual in this than in the target identification task. Specifically, for t greater than the time t_s of stimulus arrival we take $a(t) = \bar{a}(1 - e^{-r(t-t_s)})$ with $r = 50$ (time constant 0.02 sec) for target identification and $r = 10$ (time constant 0.1 sec) for the Eriksen task; also, we set $\bar{a} = 0.06$; and $\tau_c = 0.5$ sec: see Fig. 6.9(b). Additionally, we assume that t_s follows presentation of sensory cue by a processing time lag of 0.1 sec (cf. [10]). Optimal gain schedules $\bar{g}_f(t)$ for the firing rate model with these stimuli, computed as in the preceding section, are shown in Fig. 6.9(a). To produce panel (c), these gain functions were inserted into Eqn. (6.55) to yield corresponding optimal LC firing rates, the discontinuity in $\bar{g}_f(t)$ at stimulus onset having negligible effect. (Also note that assuming a smoother profile for $a(t)$ would eliminate the jump in $LC(t)$.) The similarity between overall form and decay rates of optimal gain functions $\overline{LC}(t)$ and the empirical data of Fig. 6.9(d) supports the hypothesis that the LC may affect near-optimal processing of sensory stimuli. This is true even though LC firing rates are not sustained at the initial high values that follow stimulus onset; in fact, both LC firing rate relaxation and NE time constants are compatible with optimal gain schedules.

We note that the optimal gains, and hence $LC(t)$ time courses, are computed assuming prior knowledge of the stimulus $a(t)$ and signal to noise ratio $\frac{a(t)}{c(t)}$. If this were the case, LC firing patterns should be well-correlated with stimulus onset. However, experimental data of [34], which involved variable stimulus onset times, indicates tighter correlations with behavioral responses. Here, the function $a(t)$ is perhaps better interpreted as input to motor neurons, the onsets of sensory stimuli having been detected earlier in decision layers. Thus, the most appropriate LC data for use in Fig. 6.9 would be aligned with transients in firing rates in intermediate processing layers; here we provide data aligned with sensory stimuli as the closest available surrogate. Explicit models of two-layer decision/response dynamics with variable gain are studied in the following chapter.

6.6   Discussion and conclusions

In this chapter we explicitly compute optimal gain trajectories for one-dimensional, linearized reductions of simplified models for competing neural groups involved in decisions between two alternatives. We first develop a piecewise linear approximation to the canonical sigmoidal activation or firing rate function. The resulting two-dimensional piecewise linear SDEs (6.10-6.11) introduced in Section 6.3.3 form a

midpoint in our simplification process. This system can be easily solved on each of nine 'tiles' forming its phase plane, but solutions must be assembled by matching constants of integration. To illustrate this, we focus on two specific cases in Section 6.3.4, motivated by the moving dots' paradigm [18, 156, 72], that correspond to distinct stimulus presentation conditions and rely on different neural mechanisms to implement transient effective gain values.

In case 1, the development of salience (i.e., $a_1 \neq a_2$), in sensory stimuli at time t_s is not accompanied by large changes in the stimulus magnitudes; in fact the summed magnitude is unchanged. This mild stimulus onset is insufficient to move solutions between tiles, so variations in gain must result from modulation of the neural activation function itself, presumably via influence of other brain areas. However, in case 2, the appearance of salience is accompanied by large changes in stimulus magnitude, either due to properties of the stimulus itself or due to additive biases that shift the activation function to the left, as has been proposed by connectionist models that address the effects of attention [125, 39]. In this case, no external modulation is required, since the decision dynamics themselves move the system between regions of the activation function where desired sensitivities (and hence gains) are achieved. The possibility that neural systems are tuned so that the presence of target stimuli causes solutions to move into sensitive regions of their activation functions has been previously suggested in behavioral neuroscience [154]; here we reformulate this idea in terms of optimal signal processing.

We end by showing that the (non-unique) optimal gain schedules for the firing rate model include time courses that are consistent with release of norepinephrine due to transient increases in the activity of neurons in locus coeruleus.

The external modification of gain considered in case 1 assumes prior knowledge of the time course of the absolute values of sensory inputs $a_j(t)$, the task of the decision maker being merely to identify their signs. In the following chapter the more general case is treated in which this information is not available, and strategies must additionally include a mechanism for detecting increases in the signal to noise ratio of sensory inputs.

Chapter 7

Adaptive optimization of decision tasks via the locus coeruleus

7.1 Chapter outline

In this chapter, we assess the extent to which simplified time-dependent gain schedules, subject to the constraints that are likely to affect the locus-coeruleus in certain decision tasks, can optimize performance. In particular, we abandon the assumption of the previous chapter that the gain regulation system has knowledge of the onset times of sensory stimuli. Rather, we assume, as in [167, 70], that accumulation of firing rates in decision units drives the LC. In this sense, the gain schedules considered here are adaptive and vary from trial to trial.

First, we review work on modeling decision tasks with LC-mediated adaptive gain and state the unresolved questions that are the focus of the present chapter. Then, Section 7.3.1 defines the linearized decision model used here, which is an extension of the 'case 1' firing rate model of Chapter 6. Section 7.3.2 then introduces reward rate, the figure of merit for the different decision models we study, and discusses the effect of 'premature' responses on this quantity. Thus equipped, in Section 7.3.3 we describe the method, central to the present analysis, of comparing optimal task performance of a model with LC-mediated adaptive gain against optimal task performance of a model restricted to have constant gain. Next, in Section 7.3.4, we simplify the model by eliminating redundant parameters. Having thus refined our problem, in the following Section 7.4 we present the results of numerical reward rate optimization. Next, Section 7.5 considers two modifications to the model defined in Section 7.3.1 which are motivated by the underlying neurobiology, and Section 7.6 investigates how the number of layers in a decision model affects the performance that it can achieve when optimized. We discuss the results in Section 7.7. This chapter forms the theoretical basis of a paper in progress [25].

7.2 Introduction and background

The previous chapter showed that, given knowledge of the time course of signal to noise ratios of stimuli and assuming that firing rates of competing neural pools are confined to the maximally sloped region of their activation functions (about which they may be linearized, i.e., 'Case 1' of Chapter 6), LC-mediated gain signals can optimize decision tasks. However, if such a strategy were being implemented, one would expect the time course of LC responses to be perfectly correlated with stimulus onsets from trial to trial. Figure 7.1 indicates that this is not necessarily the case: transients in LC firing rate for the two-alternative choice task data presented there are more tightly correlated with behavioral responses than stimulus presentations.

In this chapter we assess the extent to which the LC can optimize decision tasks in the absence of explicit knowledge about stimulus schedules. That is, we require that all LC-mediated gain transients are adaptively determined by the incoming signal on a trial-to-trial basis. The most general solution to this problem would involve richer signal processing strategies than are possible with the simple neural integrator models we consider; however, as detailed below, we assume a more limited capacity for adaptive gain scheduling that is based on the detection of salient sensory events. we also relate the results of this adaptive optimization study to the response-locked data of Figure 7.1.

Two previous modelling studies have addressed the effects of adaptive LC-mediated dynamic gain on decision tasks, offering intriguing results and suggesting several important questions [167, 70]. Foremost among these is the question of optimality. The work of [167, 70] showed that, by allowing greater transients in gain schedules, error rates in a simulated target detection task could be reduced while reaction times held relatively constant or also decreased. This characterizes improved cognitive performance and bypasses the classical speed accuracy tradeoff. However, to conclude that this improved performance stems from the flexibility afforded by dynamic vs. constant gain schedules, rather than being a confound of specific values of decision thresholds (held fixed as gain schedules varied in [167, 70]) or the specific gain schedules chosen for comparison, it is necessary to compare the best-possible levels of performance across the whole class of dynamic vs. constant gain schedules and over the whole range of free parameters. This approach is taken below.

Although an optimized decision model equipped with LC-mediated dynamic gain trajectories clearly will perform no worse than an equivalent optimized model for which gain values are restricted to be constant, it is not obvious that it will perform significantly better. One reason for this is that realistic gain schedules implemented in this adaptive fashion are subject to latencies: experiments by Waterhouse and collaborators have shown that increases in the sensitivity of cortical neurons to sensory inputs follow LC activation by more than 100 ms and reach a plateau at maximum effect 200-300 ms following this activation [175]. Here we model this via a delay of ≈150 ms; see below for details. Coupled with the latency from stimulus onset to triggering of the LC, gain changes therefore take effect significantly after stimuli are presented, which may minimize the benefits of these changes. Additionally, we consider only piecewise-constant gain trajectories here, which require only relatively

simple threshold-based mechanisms to implement. Further, as with any adaptive gain scheduler but with particular relevance to the class of 'jumped' gain and stimulus trajectories considered here, gain transients triggered at the wrong moments can lead to erroneous responses, possibly obviating benefits derived on correct trials. This points to an overarching 'circularity' problem that is revisited in the Discussion of Section 7.7: due to the undesirable effects of prematurely changing gain, will so much information about stimulus presence be needed in order to beneficially increase gain that the stimulus could be reliably identified (as alternative 1 or alternative 2) before gain changes typically take effect?

Another unresolved question surrounding transient gain schedules and simple cognitive tasks surrounds the design of these tasks. In [167, 70], a target detection task was studied, in which only one of two possible stimuli (the target) demands a behavioral response. For models in which transient gain changes are driven by partial evidence of the behaviorally relevant cues, this target detection task (compared with the two-alternative choice task studied here, in which both stimuli require responses) requires fewer thresholds for the triggering of gain transients and hence presents fewer opportunities for system gain to be turned up too early and evoke an erroneous response. The results below address whether performance benefits are still seen in two-alternative choice tasks.

Additionally, the role of the neural architecture with which decision tasks are solved remains to be investigated. Schall ([150] and references therein), and Reddi [141] have presented neural evidence that at least two stages of neural processing contribute significantly to variability in behavioral response times. Motivated by these studies, by the structure of the decision models of [167, 70], and by the intuitive conclusion that a flexible system mapping sensory inputs to varied motor responses (cf. [150]) must involve separate centers of sensory integration and motor processing, the present chapter extends the previous to consider two-layered decision networks. However, we also consider the effects of adaptive gain schedules on a one layer system for comparison. We also study the effects of assuming that a certain fixed level of firing in the second ('motor') layer must be achieved in order to evoke behavioral responses.

A final open question has already been introduced via Figure 7.1: at optimal performance, will transients in gain trajectories (and hence transients in LC firing rates) be more tightly locked to stimulus presentations or (simulated) behavioral responses? To the extent to which the models capture the dynamics of neural decision processes, an answer in favor of locking to responses would support the hypothesis that the LC serves to optimize decision tasks.

7.3 The RR optimization problem

7.3.1 The two layer model

Our decision model, as shown in Figure 7.2, consists of two layers, each containing two mutually inhibitory neural subpopulations whose spike rates evolve under the 'firing
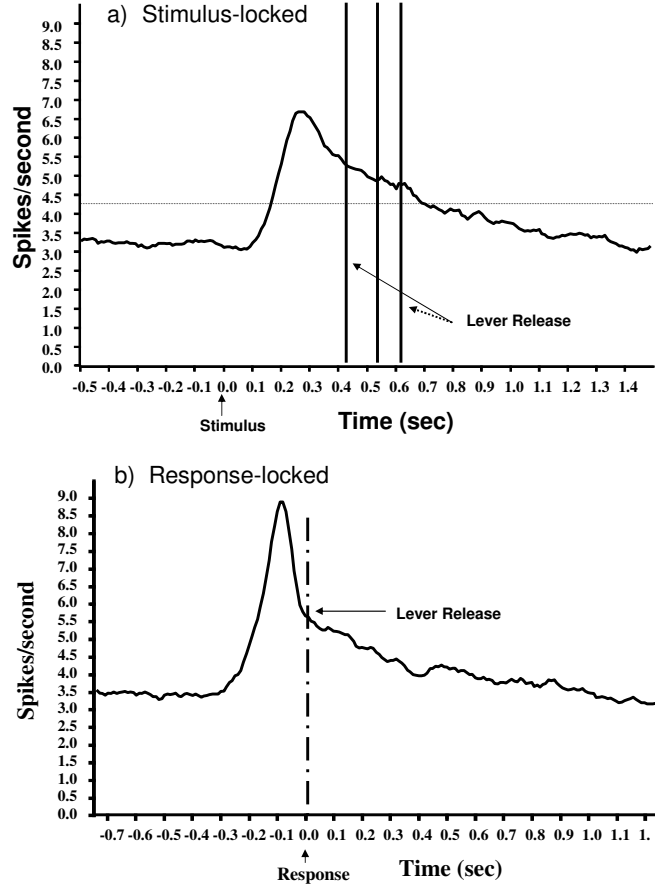
Figure 7.1: Transients in LC firing rates, here shown as peri-event time histograms (PETHs) compiled across many experimental trials of the Eriksen flanker sensory discrimination task, are more tightly correlated with behavioral responses (b) than stimulus presentations (a). Data kindly provided by the authors of [33].

rate model' of Eqns. (6.7)-(6.8). The first 'decision' layer receives sensory inputs and noise exactly as in Chapter 6:

$$\tau dy_1 = \left[-y_1 + f_{\tilde{g}_y(t)} \left(-\beta y_2 + k a_1(t) \right) \right] dt + \tilde{g}_y(t) \frac{kc\sqrt{\tau}}{\sqrt{2}} dW_t^{1'}, \qquad (7.1)$$

$$\tau dy_2 = \left[-y_2 + f_{\tilde{g}_y(t)} \left(-\beta y_1 + k a_2(t) \right) \right] dt + \tilde{g}_y(t) \frac{kc\sqrt{\tau}}{\sqrt{2}} dW_t^{2'}. \qquad (7.2)$$

The second 'response' layer receives inputs from the first (weighted by w) as well as independent noisy inputs (presumably from afferents not related to the decision task
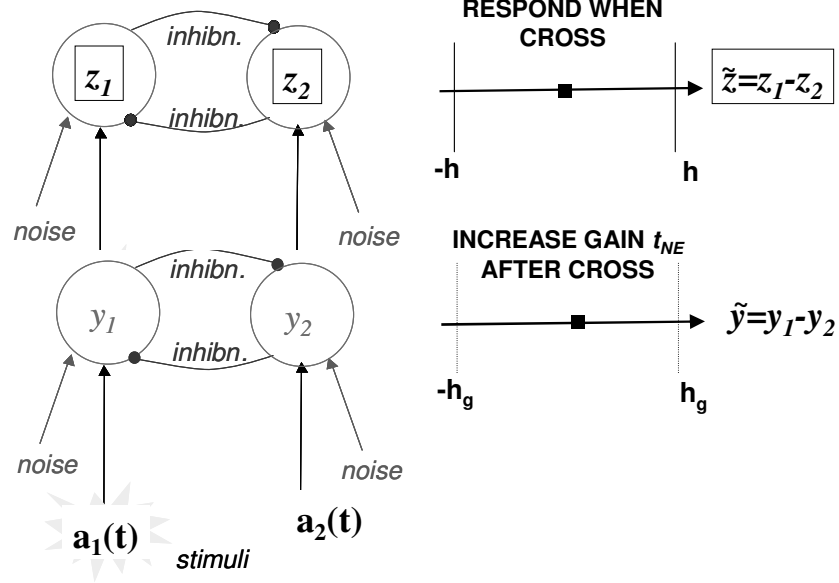
Figure 7.2: The architecture of the decision model, which consists of two layers of mutually inhibitory neural populations. The schematic to the right indicates the flow of stimulus information through the network, and the fact that, as in Chapter 6, we collapse each layer to a single variable (y or z) characterizing the difference between firing rates in the layer. When the difference $|y|$ exceeds threshold value h_g, gain levels are adjusted (following the delay τ_{NE}); when $|z|$ exceeds h, task responses are made.

being modeled):

$$\tau dz_1 = \left[-z_1 + f_{\tilde{g}_y(t)}\left(-\beta z_2 + wy_1(t)\right)\right] dt + \tilde{g}_z(t)\frac{kc\sqrt{\tau}}{\sqrt{2}}dW_t^{1''}, \qquad (7.3)$$

$$\tau dz_2 = \left[-z_2 + f_{\tilde{g}_y(t)}\left(-\beta z_1 + wy_2(t)\right)\right] dt + \tilde{g}_z(t)\frac{kc\sqrt{\tau}}{\sqrt{2}}dW_t^{2''} \qquad (7.4)$$

(the final terms in (7.1)-(7.4) are independent Wiener increments, and likewise below).

As in Section 6.3.4 (see text there for details), we define the firing rate differences $\tilde{z} = z_1 - z_2$, $\tilde{y} = y_1 - y_2$ as well as $a = a_1 - a_2$ and study the (approximate) linearized model involving only their differences:

$$\tau d\tilde{z} = \left[-\tilde{z} + \tilde{g}_z\tilde{\beta}\tilde{z} + \tilde{w}\tilde{g}_z\tilde{y}\right] dt + \tilde{g}_z kc\sqrt{\tau}dW_t^1 \qquad (7.5)$$

$$\tau d\tilde{y} = \left[-\tilde{y} + \tilde{g}_y\tilde{\beta}\tilde{y} + \tilde{g}_y ka\right] dt + \tilde{g}_y kc\sqrt{\tau}dW_t^2 \ . \qquad (7.6)$$

As already mentioned, the validity of this reduction depends on the firing rates of all populations staying in the linear regime of their activation functions; we assume that this holds throughout this Chapter, as for the 'Case 1' dynamics of Chapter 6
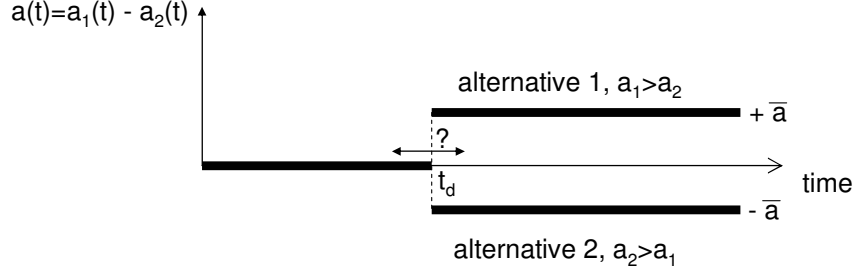
Figure 7.3: The type of stimuli considered here. Salience (i.e., the difference between strengths of inputs favoring one of the two alternatives) develops at a randomly distributed time t_d. The task of the decision maker is to determine which of the two alternative stimuli has occurred, balancing speed vs. accuracy in this discrimination in order to maximize reward rate (see text).

([20] checks the validity explicitly). Here, $\tilde{g}_z$ and $\tilde{g}_y$, respectively, are the gains of neural groups in these layers. We consider time-dependent stimuli $a(t)$ of the type illustrated in Figure 7.3: $a(t) = 0$, $t < t_d$ and $a(t) = \pm\bar{a}$, $t \geq t_d$, where t_d is the randomly distributed time of stimulus onset (see below).

Furthermore, we take initial conditions $y(0) = z(0) = 0$ at the beginning of the preparatory phase, at which time we assume firing rates are reset (e.g. by inhibitory afferents originating in the prefrontal cortex). Additionally, $\tilde{w}$ is the weight associated with inputs to the response layer from the decision layer. $k\,a$ is the rate of incoming currents associated with sensory 'evidence' of task stimuli, and $k\,c$ sets the strength of noise in afferent currents to the two populations; $k > 0$ is an arbitrary scale factor setting the magnitude of inputs to the neural populations. τ sets the timescale of neural integration, and β sets the strength of inhibition between the pair of competing neural subpopulations at each layer. The reason for the tilde notation will become clear below.

We assume that the gain values $\tilde{g}_z, \tilde{g}_y$ undergo transient changes in both layers at a (physiologically determined) delay t_{NE} following the first passage time $T_y = \inf\{t : |y(t)| > \tilde{h}_g\}$ of $\tilde{y}$ from the interval $[-\tilde{h}_g, \tilde{h}_g]$. Specifically, we assume that the gain values jump from values $\tilde{g}_z^{pre}, \tilde{g}_y^{pre}$ to $\tilde{g}_z^{post}, \tilde{g}_y^{post}$ at this time, and we refer to $\tilde{h}_g$ as the 'gain threshold.' This mechanism for adjusting gain generalizes the approach of [167, 70] from the target identification task (in which gain transients are driven only by firing rates in the population responsive to target stimuli) to the two-alternative forced choice task (in which gain transients may be driven by the population responsive to either of the two possible task stimuli). We assume that NE levels have decayed to their baseline levels $\tilde{g}_z^{pre}, \tilde{g}_y^{pre}$ by the time each subsequent trial begins.

A response corresponding to alternative 2 (i.e., that "$a < 0$") is made at the passage time $T_z = \inf\{t : |\tilde{z}(t)| > \tilde{h}\}$ if $\tilde{z}$ first exits the interval $[-\tilde{h}, \tilde{h}]$ by crossing the barrier $-\tilde{h}$ (i.e., $\tilde{z}(T_z) = -\tilde{h}$), and vice-versa for alternative 1 ($a > 0$) and the barrier at $\tilde{h}$; we call $\tilde{h}$ the response threshold. To determine optimal performance of
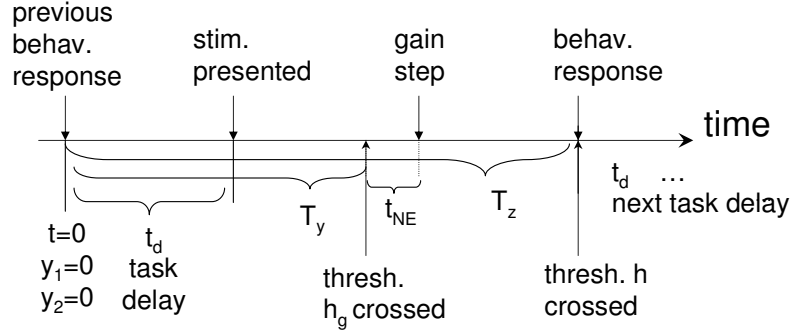
Figure 7.4: Schematic of the timing of the model decision task. Following a behavioral response on the 'previous' trial, model time t is reset to $t = 0$, initial conditions $y_1(0) = y_2(0) = 0$ applied, and firing rates begin evolving under Eqns. (7.5)-(7.6). The sensory stimulus is presented following a (randomly distributed) delay t_d. At a delay t_{NE} following the first passage time T_y of the firing rate difference y across thresholds $\pm h_g$, model gains undergo a step change. Finally, the behavioral response on the 'current' trial is made at time T_z. Although the ordering of t_d, T_y, and T_z displayed here is typical, it is not enforced: for example, on some trials gain threshold crossings at T_y could occur before stimulus presentations at t_d.

the models under different architectures and assumptions about the time dependence of gain, we shall allow gains $\tilde{g}_y$ and $\tilde{g}_z$ and thresholds $\tilde{h}$ and $\tilde{h}_g$ to vary freely, as specified in greater detail below.

7.3.2 Task setup and reward rate

We model a task in which the objective is to correctly identify which of two alternative stimuli have been presented on each trial so as to maximize the reward rate (RR) [72], or the rate at which correct responses are made. Trials occur in a long sequence, with randomized delays between each behavioral response and the presentation of the subsequent stimulus. We assume that these stimulus presentations are accompanied by a step in signal coherence a entering the first layer, as in Section 6.3.4 (cf. Figure 7.3) but with variable onset times t_d. We assume that no other changes in signal coherence a occur. To make contact with typical behavioral experiments, we take t_d to be uniformly distributed between 1 and 3 seconds for all of the results reported below.

In defining RR, we must specify how 'premature' responses made in the period $[0, t_d]$ (that is, between the previous response and the presentation of the current sensory cue) are scored as correct vs. incorrect. Furthermore, we must determine what effect, if any, these responses will have on the delay before presentation of the next sensory cue. We adopt the following protocol, based on task designs in common use (including in [33]): all premature responses are counted as errors, and the next trials are started immediately. In this case, we relate the reward rate RR to random

variables characterizing the decision process as follows:

$$RR = \frac{\mathbb{E}C(\tilde{z}(T_z))}{\mathbb{E}(T_z)} \quad , \tag{7.7}$$

where the function $C(\tilde{z}(T_z))$ measures the correctness of the firing rate difference $\tilde{z}(T_z)$ at the response time T_z by taking value 1 (resp. 0) if, on a trial, $a > 0$, $\tilde{z}(T_z) = \tilde{h}$, and $T_z \geq t_d$ (resp. $a < 0$, $\tilde{z}(T_z) = -\tilde{h}$, and $T_z \geq t_d$). See 7.4 for a schematic of the various times and delays that contribute to the RR. Thus defined, RR is the expected fraction of correct responses divided by the expected time elapsed for each response. RR is therefore the rate at which correct responses are made and has the units 1/time. We note that RR cannot be written as an expectation of a single function of firing rates at the hitting time T_z (unlike other objective functions typically used in sequential analysis, such as Bayes risk [114]), but rather takes quotient form. The RR measure adopted here is, however, directly related to the cognitive tasks we consider, in which participants are typically instructed to make as many correct responses as possible in a fixed interval.

Finally, we note that simulations were also performed with ranges of t_d other than uniform distribution from 1 to 3 seconds described above (including fixing t_d to take only a single value, nonetheless assumed unknown to the decision maker) and other protocols for treating premature responses (such as simply ignoring any threshold crossing prior to stimulus presentation time t_d at which responses are immediately recorded if the firing rate difference $|z|$ remains above threshold). In these cases, the qualitative relationship between optimal performance of the constant vs. adaptive gain models that we next describe is preserved.

7.3.3 Optimizing decision models under adaptive vs. constant gain schedules

We now state the optimization problem of tuning model parameters to achieve maximal reward rate using LC-mediated adaptive gain:

$$RR_d = \max \ RR(\tilde{g}_y^{pre}, \tilde{g}_z^{pre}, \tilde{g}_y^{post}, \tilde{g}_z^{post}, \tilde{h}, \tilde{h}_g) \text{ under Eqns. (7.5)-(7.6)} \ . \tag{7.8}$$

The subscript d indicates that gain is 'dynamic:' allowed to adaptively change as described above.

Next, we determine the extent to which LC-mediated adaptive gain contributes to reward rates found by solving (7.8). As mentioned in the Introduction, we accomplish this by simply asking what the best performance would be in the absence of such an adaptive gain mechanism (but allowing freedom in choosing all other parameters). That is, we determine the optimal reward rate RR_c for constant gain schedules by restricting $\Delta g = 0$:

$$RR_c = \max \ RR(g_y^{pre}, g_z^{pre}, h, h_g) \text{ under Eqns. (7.5)-(7.6) with } \Delta g = 0 \ . \tag{7.9}$$

In Sections 7.4-7.6 below, we directly compare (refined versions of) RR_d and RR_c

under various assumptions on task protocols and model architecture.

7.3.4 Eliminating redundant parameters

We now eliminate redundant parameters in the model to clarify the assumptions implicit in the model and hence the generality of the results that derive from optimizing it. First we show that the scale factor k (which sets the absolute magnitude of stimulus and noise inputs but not their ratio) may be eliminated by rescaling thresholds $\tilde{h}$ and $\tilde{h}_g$. Dividing (7.5)-(7.6) through by k and redefining $\tilde{\tilde{z}} = \tilde{z}/k$ and $\tilde{\tilde{y}} = \tilde{y}/k$ we obtain

$$\tau d\tilde{\tilde{z}} = \left[-\tilde{\tilde{z}} + \tilde{g}_z \tilde{\beta}\tilde{\tilde{z}} + \tilde{w}\tilde{g}_z\tilde{\tilde{y}} \right] dt + \tilde{g}_z c\sqrt{\tau}dW_t^1 \tag{7.10}$$

$$\tau d\tilde{\tilde{y}} = \left[-\tilde{\tilde{y}} + \tilde{g}_y \tilde{\beta}\tilde{\tilde{y}} + \tilde{g}_y a \right] dt + \tilde{g}_y c\sqrt{\tau}dW_t^2 \ . \tag{7.11}$$

After scaling thresholds $\tilde{h}$ and $\tilde{h}_g$ by the (positive) factor $1/k$ (matching the scaling of variables $\tilde{z}$, $\tilde{y}$), we obtain exactly the same statistics for (7.10)-(7.11) as for the original system (7.5)-(7.6); furthermore, since thresholds are free parameters in the optimization problem (7.8), we have lost no generality in eliminating the parameter k, and see that only the signal to noise ratio a/c affects optimal reward rates.

The rate a has units of $1/t$, and c has units of $1/\sqrt{t}$. We choose units of time to be seconds, and fix the ratio $(\bar{a}/c)^2 = 8\ sec.^{-1}$, a value derived from maximum likelihood fits of reaction time distributions from two-alternative choice task experiments in [13]. Furthermore, we fix $\tau = 1$, noting that the results we derive from the model with $\tau = 1$ correspond to results from $\tau \neq 1$ and a suitably rescaled signal to noise ratio (following division of (7.10)-(7.11) by τ and another rescaling of the firing rate variables).

Additionally making the definitions $g_y = \tilde{\beta}\tilde{g}_y$, $g_z = \tilde{\beta}\tilde{g}_z$, $y = \tilde{\beta}\tilde{\tilde{y}}$, $z = \tilde{\beta}\tilde{\tilde{z}}/w$, and $w = \tilde{w}/\tilde{\beta}$, (7.10)-(7.11) become

$$dz = \left[-z + g_z z + g_z y \right] dt + g_z \frac{c}{w}dW_t^1 \tag{7.12}$$

$$dy = \left[-y + g_y y + g_y a \right] dt + g_y c dW_t^2 \ . \tag{7.13}$$

Recall that the variables $\tilde{g}_y$ and $\tilde{g}_z$ can take arbitrary (positive) values in the optimization scheme, so that rescaling them by the constant $\tilde{\beta}$ to form g_z, g_y has no effect on the optimization problem. Furthermore, since the thresholds $\tilde{h}$, $\tilde{h}_g$ are also free variables, the rescaling of variables y, z is again without consequence for the optimization problem: the rescaled system (7.12)-(7.13) with free parameters g_z, g_y as well as thresholds h_g and h produces the same optimal error rates and reaction times (and hence reward rates) as the original system (7.5)-(7.6) with free parameters $\tilde{g}_z$, $\tilde{g}_y$, $\tilde{h}_g$ and $\tilde{h}$.

The only parameter in (7.12)-(7.13) that remains to be defined is the weight w, which we set as follows. Recall that the variable z represents firing rates in the 'response' cortical populations which drive the motor neurons enacting task decisions.

In any flexible neural architecture these cortical populations receive inputs not only from the decision layer populations responsive in the specific two-alternative task modelled here, but also from a diverse set of other neurons and brain areas. It is fluctuations in these inputs that are modelled by the input noise $\frac{c}{w}dW_t^1$. The magnitude of (mean values of) y, which is the 'signal' component of the input to the response layer, are determined by a: thus, the signal to noise ratio for this layer is $\approx aw/c$. For simplicity, we set $w = 1$ so that signal to noise ratios in both layers are the same. Therefore, the final system of equations is

$$
\begin{aligned}
dz &= [-z + g_z z + g_z y]\, dt + g_z c dW_t^1 \tag{7.14} \\
dy &= [-y + g_y y + g_y a]\, dt + g_y c dW_t^2 \ , \tag{7.15}
\end{aligned}
$$

and we study the optimization problems equivalent to (7.8), (7.9):

$$
RR_d = \max\ RR(g_y^{pre}, g_z^{pre}, g_y^{post}, g_z^{post}, h, h_g) \text{ under Eqns. (7.14)-(7.15)}\ , \tag{7.16}
$$

$$
RR_c = \max\ RR(g_y^{pre}, g_z^{pre}, h, h_g) \text{ under Eqns. (7.14)-(7.15) with } \Delta g = 0\ . \tag{7.17}
$$

The extent to which norepinephrine adjusts the gain of a particular cortical population depends on the density and type of norepinephrine receptors in that population (as well as, e.g., connectivity within and afferent inputs to that population) [175]. However, for simplicity we assume here and below that the phasic LC impulse results in gain changes of the same magnitude in both the first and second layers. That is, a delay t_{NE} following the time T_y at which thresholds are crossed in the first layer, $g_y^{pre} \to g_y^{pre} + \Delta g$ and $g_z^{pre} \to g_z^{pre} + \Delta g$. This reduces by one the number of free parameters in the optimization problem (but does not otherwise affect the generality of the results with regard to redundant parameters), giving the modified dynamic gain optimization problem:

$$
RR_d = \max\ RR(g_y^{pre}, g_z^{pre}, \Delta g, h, h_g) \text{ under Eqns. (7.14)-(7.15)}\ . \tag{7.18}
$$

Here, RR_d denotes the best possible reward rate that can be achieved with this adaptive gain scheme. Simulations indicate that the problems (7.18) and the slightly more general (7.16) give the same optimal RR's within $\sim 2\%$.

7.4 Numerical optimization

7.4.1 Algorithm

We use the SUBPLEX optimization algorithm [147], a generalization of the Nelder-Mead Simplex method well suited for noisy objective functions, to solve the problem of optimizing RR under various conditions. The evaluation of RR (7.7) for at each set of parameters sampled by the algorithm is done by Monte-Carlo simulation of 200,000 'trials' of the SDEs (7.14)-(7.15), each with simulated inter-trial interval t_d (i.e. 'prepcycle' duration) drawn from a uniform distribution between 1 and 3 seconds.
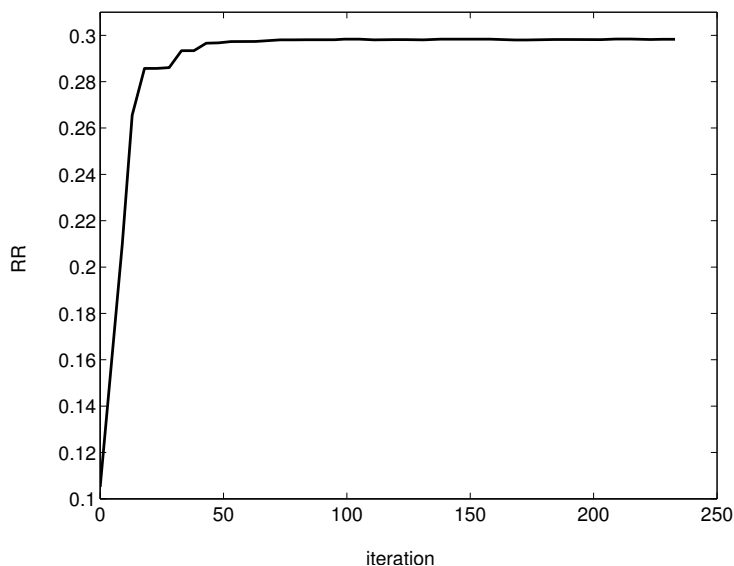
Figure 7.5: Typical convergence of the SUBPLEX algorithm on parameters yielding an (approximately) optimal RR.

The standard deviation in RR for a typical evaluation (at fixed parameter values) with this number of trials is $0.0003 \ sec^{-1}$. Convergence of the SUBPLEX algorithm for a randomly chosen set of initial conditions is shown in Figure 7.5.

7.4.2 The standard parameter set

We now collect the values of model parameters that we hold fixed in the evaluation of RR via (7.14)-(7.15). These values, motivated in the text above, are: $\bar{a} = 2 \ sec^{-1}$, $c = 1/\sqrt{2} \ sec^{-1/2}$, and delays t_d between behavioral responses and presentation of the next sensory cues are uniformly distributed between 1 and 3 sec. All of these quantities may be derived from task design and behavioral data under certain assumptions (see above).

7.4.3 Range of reward rate values

It will be instructive to study how the reward rates achieved by the (linearized) two-layer model (7.14)-(7.15), with or without dynamic gain, compare with the reward rate achievable by an optimal decision maker. As discussed in Section 6.4.1, this optimal decision maker performs the sequential probability ratio test, which assumes knowledge of the time course of signal-to-noise ratios. From the results in Chapter 6, a one-layer linearized network can perform the SPRT if it is provided with perfect apriori knowledge of stimulus onset times and the capacity to instantaneously adjust gain to exploit this knowledge. Again drawing on Chapter 6, the optimal gain for the firing rate model of (6.16) for the piecewise-constant stimuli considered here is: $g(t) = 0, t < t_d$, $g(t) = 1, t \geq t_d$. In this case error rate and reaction time depend on

145

the model parameters as follows ([58, 13]):

$$ER = \frac{1}{1 + \exp(2h\bar{a}/c^2)} \tag{7.19}$$

$$RT = \frac{h}{\bar{a}} \tanh\left(\frac{h\bar{a}}{c^2}\right) \quad . \tag{7.20}$$

As shown in [13], the unique optimal value of h optimizing

$$RR = \frac{1 - ER}{RT + \mathbb{E}(t_d)} \tag{7.21}$$

is given by the solution to

$$\exp(2h\bar{a}/c^2) - 1 = 2\beta(\mathbb{E}(t_d) - h/\bar{a}) \quad . \tag{7.22}$$

Solving (7.22) using Newton's method (for the standard parameter set, that is, for $\mathbb{E}(t_d) = 2$, $\bar{a} = 2$, $c = 1/\sqrt{2}$), we obtain $h = 0.212$. Using this to find ER and RT and then inserting these into (7.21) gives the 'ceiling' value $RR_{ceil} = 0.440\ sec^{-1}$ achieved by the SPRT with optimal thresholds.

We next compute a complementary floor value for RR so that we can define a suitable 'range' of RR as $[RR_{floor}, RR_{ceil}]$. We let the floor value be achieved by chance guessing, assuming that a decision maker employing this strategy chooses hypothesis 0 or 1 without paying any attention to task stimuli, at times separated from the previous response by a uniform distribution between 0 and 3 seconds (i.e., up to the maximum possible value of t_d. Since $t_d > 1$ sec., a third of these responses will occur before the stimulus has been presented, and hence are guaranteed to result in errors. The rest of the responses will be correct 50% of the time. Thus, $1 - ER = \frac{1}{3} * 0 + \frac{2}{3} * \frac{1}{2} = \frac{1}{3}$. The mean elapsed time between responses using this strategy is $\frac{3}{2}$ sec. Thus the floor value produced is $RR_{floor} = \frac{1}{3} * \frac{2}{3} \approx 0.222\ sec^{-1}$. In conclusion, the range of RR is $[RR_{floor}, RR_{ceil}] = [.222, .440]$ for the standard parameter set.

7.4.4 Optimal reward rates for the 'standard' parameter set

The results of the optimization problems (7.18)-(7.9) for the standard parameter set are shown in Figure 7.6. The best reward for the dynamic gain problem (7.18) is $RR_d = 0.299\ sec.^{-1}$ while $RR_c = 0.267\ sec.^{-1}$ is achieved for the constant gain version (7.9). Thus, transient modulation of the decision process enabled by the LC allows the decision maker to enhance rewards by 12 percent (on average) during the interval over which the task is performed.

We now view these results in terms of the range of RR values derived above in Section 7.4.3. Letting $RR = RR_{floor}$ be 0% through the range and $RR = RR_{ceil}$ be 100% through the range, the optimized two-layer constant gain model achieves a RR 20.6% of the way through the range; this improves to 35.3% with the addition of dynamic gain.

As Figure 7.6 indicates, the SUBPLEX algorithm converged to slightly different
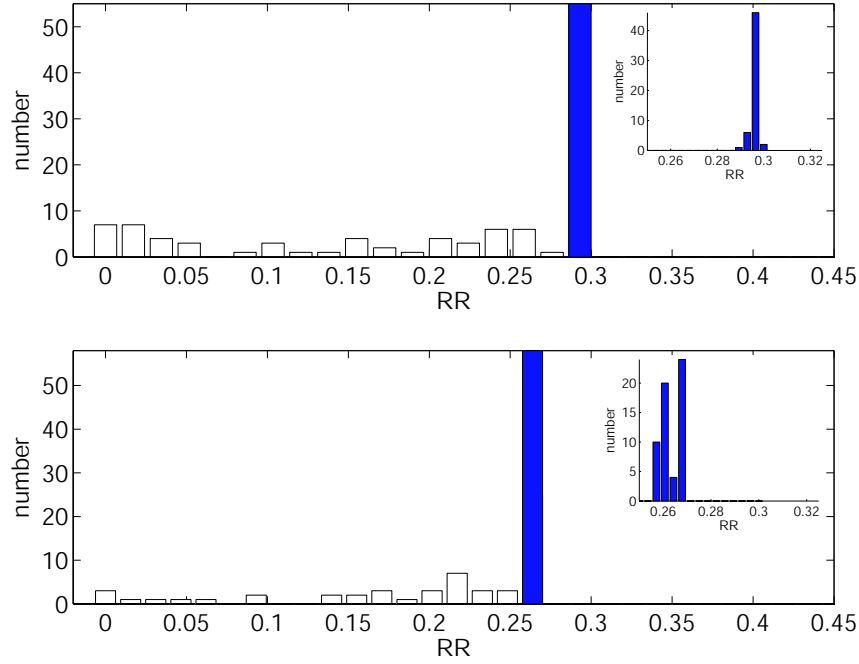
Figure 7.6: Filled bars–histogram of reward rates found using the SUBPLEX optimization algorithm; the vertical axis measures the number of runs of the SUBPLEX algorithm (each with a different, randomly chosen set of initial parameter values) that converged to particular RR values. (top) for the dynamic, adaptive gain optimization problem (7.16); (bottom) for the constant gain optimization problem (7.17). Insets give zoomed view with smaller histogram bin size. As reported in Section 7.4.4, the maximal reward rate obtained via SUBPLEX optimization for the adaptive gain case was RR_d=0.299 $sec.^{-1}$ (rightmost filled bar, top insert) and the maximal reward rate for the constant gain case was RR_c=0.267 $sec.^{-1}$ (rightmost filled bar, bottom insert). Outlined bars–for comparison, histogram of reward rates for randomly chosen, non-optimized parameter values. As expected, reward rates are much lower for these parameters.

RR values each time it was run (with randomly sampled initial values for all free parameters). Figure 7.7 shows the values of two of the free parameters following optimization for the dynamic gain optimization problem (7.18). There is a family of different g_y^{pre} and h_g pairs (for example) that give similar RR values (the RR's reported in this paper are the highest produced by any parameter values belonging to this family). Values of h_g depend monotonically on g_y^{pre} in this 'nearly optimal' regime, preserving the extent of evidence in favor of one or the other hypothesis necessary to trigger an increase in gain. Figure 7.8 demonstrates the analogous relationship between h_g and g_z^{pre} for the constant gain problem (7.9).

Figure 7.7: Values of two of the parameters to which the optimization algorithm converges when solving the dynamic gain optimization problem (7.18). Initial values of the parameters for each randomly initialized run of the algorithm lie at one end of the dotted lines, with final values indicated by the dots at the other end. The black square indicates the parameter values for the best-obtained RR.

7.4.5 Predicted behavioral data for the standard parameter set

We now describe the behavioral statistics that characterize optimized performance for two layer models with dynamic gain. These statistics were produced for the parameter set giving the best RR among all parameter sets to which the SUBPLEX algorithm converged (corresponding to the black square in Figure 7.7): $g_y^{pre} = 0.873$, $g_z^{pre} = 0.474$, $\Delta g = 3.33$, $h_g = 1.43$, and $h = 1.86$.

Typical trajectories for the firing rate equations (7.14) -(7.15) with these optimal parameters are shown in Fig. 7.9. Note that, due to the relatively large value of Δg, (7.14)-(7.15) become strongly unstable Ornstein-Uhlenbeck processes following the jump in gain (i.e., a delay t_{NE} after the thresholds $\pm h_g$ are crossed). Figure 7.10 displays the reaction time distribution (stimulus-locked times $T_z - t_d$) resulting from an ensemble of such trials, 16.8% of which resulted in premature responses and 2.0% of which were errors made following stimulus presentation. Recall that the decision model assumes no explicit or implicit cue indicating stimulus onset, leading to the high level of premature responses at optimal performance.

Finally, Figure 7.11 demonstrates that the times T_y at which the thresholds $\pm h_g$ were crossed are more tightly correlated with behavioral response than with stimulus onset times. According to the model of LC-mediated adaptive gain adopted here, the LC commences a period of increased firing at these threshold crossing times (which are followed t_{NE} later by the gain increase Δg). Thus, the results of Figure 7.11 may be interpreted as histograms of predicted onset times for bursts of LC activity. These
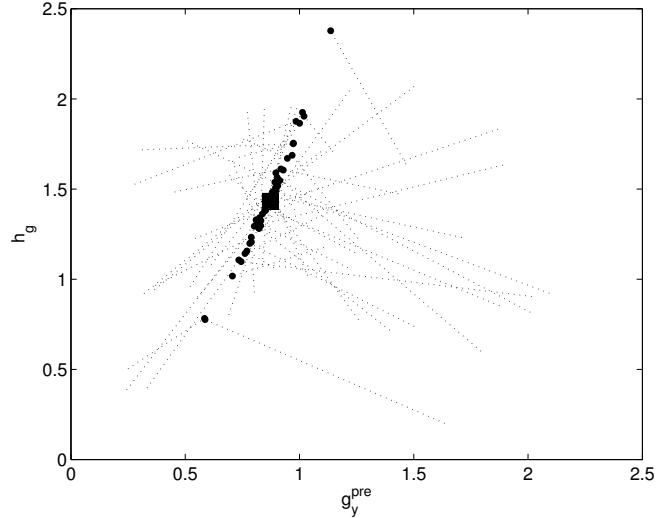
148
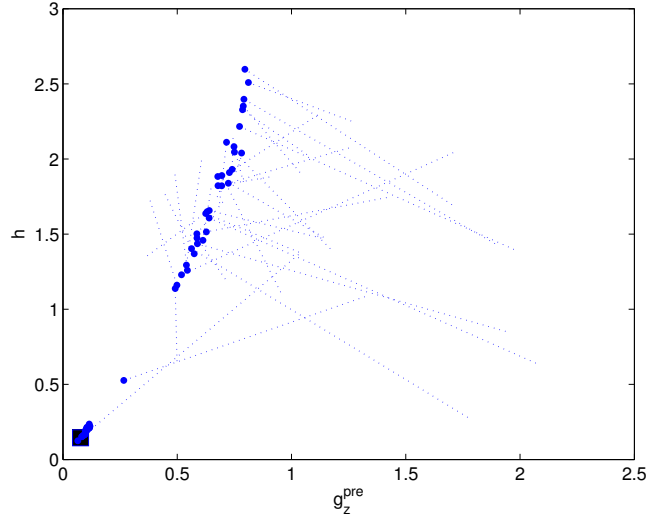
Figure 7.8: Values of two of the parameters to which the optimization algorithm converges when solving the constant gain optimization problem (7.9). Initial values of the parameters for each randomly initialized run of the algorithm lie at one end of the dotted lines, with final values indicated by the dots at the other end. The black square indicates the parameter values for the best-obtained RR.

histograms, when locked to simulated behavioral response times T_z vs. stimulus onset times t_d, display the same qualitative narrow vs. broad trend as the experimental histograms of trial-averaged LC firing rates displayed in Fig. 7.1.

7.5 Two biologically motivated constraints to the standard parameter set

7.5.1 Fixed motor thresholds

Next we define a related optimization problem motivated by a plausible neurobiological constraint and suggested by Josh Gold (University of Pennsylvania [74]). We take a more conservative view of optimization in neural decision networks by assuming that the motor threshold h is not available for adjustment from task-to-task, but is rather fixed at some sufficiently high value so as to avoid being prematurely crossed in any of a variety of different tasks with a varying signal-to-noise ratios and stimulus magnitudes. Here, we fix $h = 5$, roughly 2-3 times typical gain values for nearly-optimized solutions for the problem (7.18) in which h was allowed to vary freely. That is, we numerically solve the constrained optimization problem

$$RR_d = \max \ RR(g_y^{pre}, g_z^{pre}, \Delta g, h_g) \text{ under Eqns. (7.14)-(7.15) with } h = 5 \ . \quad (7.23)$$
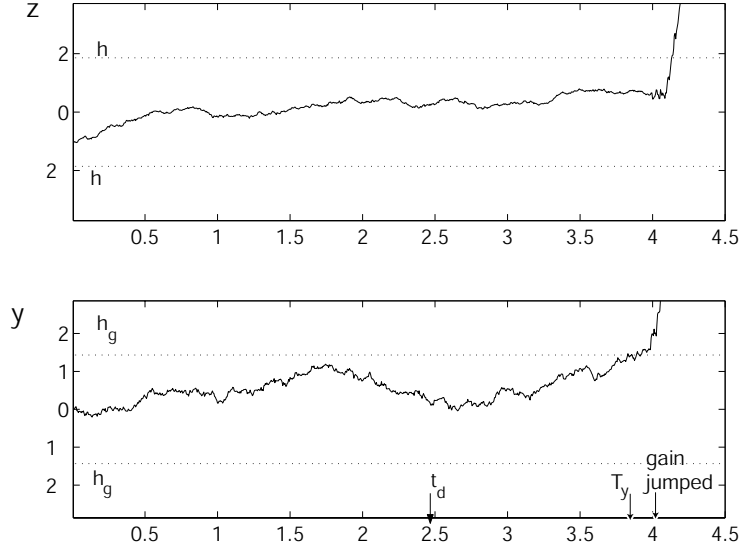
Figure 7.9: Typical trajectories for the optimized two-layer decision model (7.14)-(7.15) with dynamic gain.

for dynamic gain and

$$RR_c = \max\ RR(g_y^{pre}, g_z^{pre}, h_g) \text{ under Eqns. (7.14)-(7.15) with } \Delta g = 0 \text{ and } h = 5\ .$$
(7.24)

for the constant gain case.

The results of the optimization problems (7.23)-(7.24) in this case are as follows: the best RR for the dynamic gain problem (7.23) is RR_d=0.299 $sec.^{-1}$ while RR_c=0.247 $sec.^{-1}$ is achieved for the constant gain version (7.24). Thus, transient modulation of the decision process enabled by the LC allows the decision maker to enhance expected rewards by 21 percent over any fixed task interval. Furthermore, note that the RR value 0.299 $sec.^{-1}$ obtained for (7.23) matches that obtained for the problem (7.18), which allowed h to vary freely. The conclusion is that fixed motor thresholds have an insignificant effect on the reward rate that a system with dynamic gain can achieve, but are detrimental when gain is constant.

In terms of their position in the dynamic range $[RR_{floor}, RR_{ceil}]$, the optimized two-layer constant gain model achieves a RR level 11.3% of the way through the range; this improves to 35.3% with the addition of dynamic gain.

7.5.2 Stimulus-dependent noise

Next we allow for depressed noise in inputs to the first layer before the stimulus coherence emerges at time t_d, presuming that additional fluctuations in afferent inputs accompany the appearance of this coherence. Specifically, we set $c = \frac{1}{4} * \frac{1}{\sqrt{2}}$ for $t < t_d$ and $c = \frac{1}{\sqrt{2}}$ for $t \geq t_d$. The $\frac{1}{4}$ is an arbitrary value chosen to illustrate the present situation.
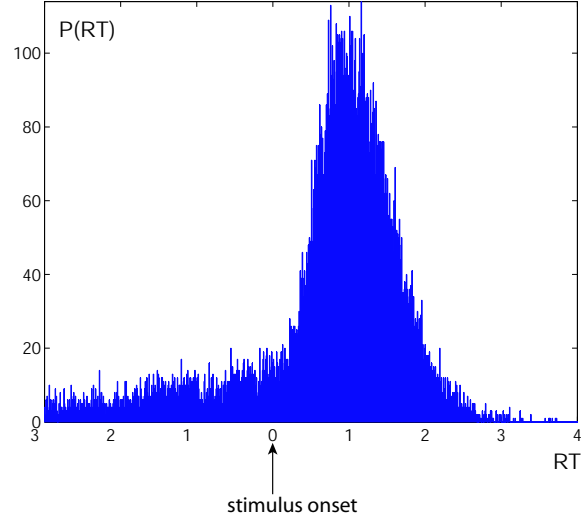
Figure 7.10: Distribution of reaction times for the optimized two-layer decision model with dynamic gain, relative to stimulus presentation; negative values correspond to premature responses.

The optimal reward for the dynamic gain problem (7.18) in this case is RR_d=0.361 $sec.^{-1}$. For comparison, RR_c=0.311 $sec.^{-1}$ was obtained achieved for the constant gain problem (7.9). Here, then, the LC enables reward rate improvements of 16 percent over any fixed task interval. Measured relative to the dynamic range, the optimized two-layer constant gain model achieves a RR level 40.1%, improving to 63.7% under dynamic gain.

7.6 One layer decision models

In this section we investigate the role that the two-layer architecture of the model has in determining the processing benefits of LC-mediated transient gain over the 'control' hypothesis of constant gain. We collapse the two-layer model of Fig. 7.2 to an effectively single-layer decision model by assuming that both gain effects and behavioral responses are driven by threshold crossing in the first layer. In other words, gain changes continue to be triggered exactly as above, but responses are made at the passage time $T = \inf\{t : |\tilde{y}(t)| > \tilde{h}\}$. See Fig. 7.12.

The optimization problem for the one layer model is

$$RR_d = \max \ RR(g_y^{pre}, \Delta g, h_g, h) \text{ under Eqn. (7.15)} \ . \tag{7.25}$$

for dynamic gain and

$$RR_c = \max \ RR(g_y^{pre}, h) \text{ under Eqn. (7.15)} \ . \tag{7.26}$$

for the constant gain case. Solving (7.25)-(7.26) via numerical optimization, we find
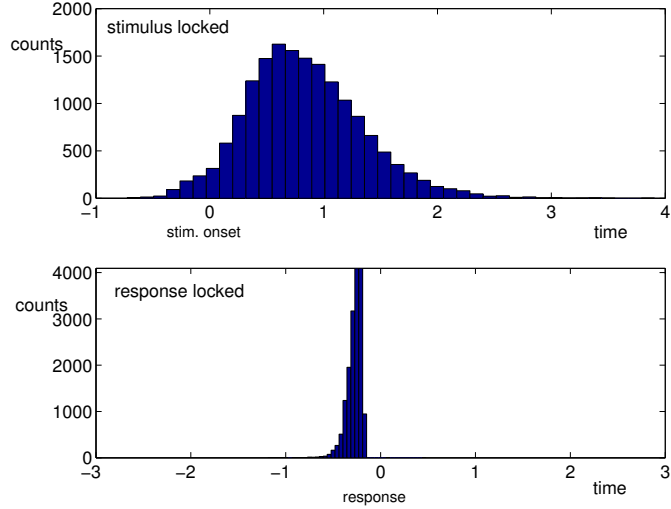
Figure 7.11: Histograms of latencies between first passage times T_y –times at which the LC fires its phasic burst in the model– and (top) stimulus onset times t_d vs. (bottom) response times T_z, all for the optimized two-layer decision model with dynamic gain.

that the best reward for the dynamic gain version (7.25) is RR_d=0.339 $sec.^{-1}$ while RR_c=0.337 $sec.^{-1}$ is achieved for the constant gain version (7.9): a negligible difference. Therefore the model of LC-mediated adaptive gain has no effect on performance for a one-layer decision maker. The reward values thus obtained by the one-layer model are approximately 53% of the way through the RR range, a significant improvement over all two-layer models. We will revisit this fact in the Discussion of Section 7.7.

We also studied versions of (7.25)-(7.26) with fixed motor threshold $h = 5$ (as in Section 7.5.1 above), with the same threshold crossing effects as in Fig. 7.12. In this case the optimal performance for the adaptive-gain model is RR_d=0.306 $sec.^{-1}$, while RR_c=0.281 $sec.^{-1}$ for constant gain. Thus, when motor thresholds are fixed, adaptive gain scheduling yields advantage for the one-layer model. Furthermore, unlike for the two-layer model, optimal RR values are reduced by fixing the motor threshold.

7.7   Discussion

This chapter shows that adaptive gain changes, potentially mediated by the locus coeruleus (LC), can help optimize performance on simulated sensory discrimination tasks, even when no knowledge of stimulus timing is assumed. This type of task is complementary to that studied in Chapter 6, where the time course of stimuli was either assumed to be known explicitly ('case 1' of Section 6.3.4) or to be expressed implicitly via strong onset effects, which evoke stimulus-locked changes in the processing of inputs ('case 2' of Section 6.3.4).

The primary model considered in this Chapter has two layers: the first integrates

Figure 7.12: The architecture of the one-layer decision model. Thresholds for both gain adjustments and behavioral responses are collapsed to a single layer. The question mark indicates that the gain thresholds h_g are not apriori restricted to lie within the response thresholds h, and may take values outside these thresholds (hence ensuring that gain will be constant) for optimal performance.

sensory input directly and the second accumulates this filtered input (as well as noise from other brain areas) and translates it into motor response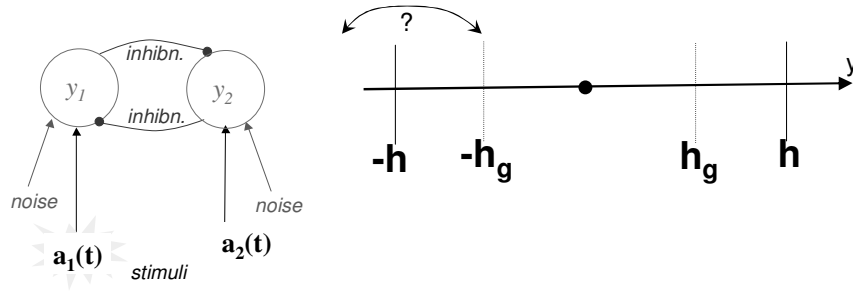s via threshold crossing events. Gain transients occur after a physiologically motivated delay t_{NE} following threshold crossings in the first layer − in this sense, gain schedules adapt 'online' according to accumulated sensory information. The linearization that we adopt allows a clear understanding of parameter effects, which we use to obtain a simpler set of optimization problems without sacrificing the generality of their solutions. By comparing optimal model reward rates in the presence of simulated LC mediated gain changes with the (separately) optimized reward rates in the absence of such gain changes, we determine the extent to which these gain changes contribute to enhanced task performance, all for a 'standard parameter set' derived from fits to experiments. The results, summarized in the bottom rows of Figures 7.13-7.14, indicate that a modest but significant improvement in reward is attributable to the LC mediated gain mechanism. Additionally, the statistical variations in the optimal model gain transients from trial-to-trial agree with trends reported in recent experimental studies involving direct recordings from the LC [33] (see Figure 7.1). This provides converging evidence for the hypothesis of [167] that the LC plays a part in optimizing the dynamics of simple decision tasks.

The improvement in reward rate that LC-mediated gain transients lend to two-layer models does not carry over to one-layer models (Figure 7.12), for which the adaptive gain transients considered here do not always yield reward rates higher than those attainable with constant values of gain (top row of Figure 7.13). Intuitively, this is because the first layer directly integrates incoming sensory inputs, and once sufficient evidence has accumulated in it to determine the presence of a stimulus and hence the advantage of a concurrent increase in gain, there is also sufficient evidence to identify which of the two possible stimuli was actually presented. Thus, optimal strategies directly produce a behavioral response at this point, instead of relying on a gain transient.
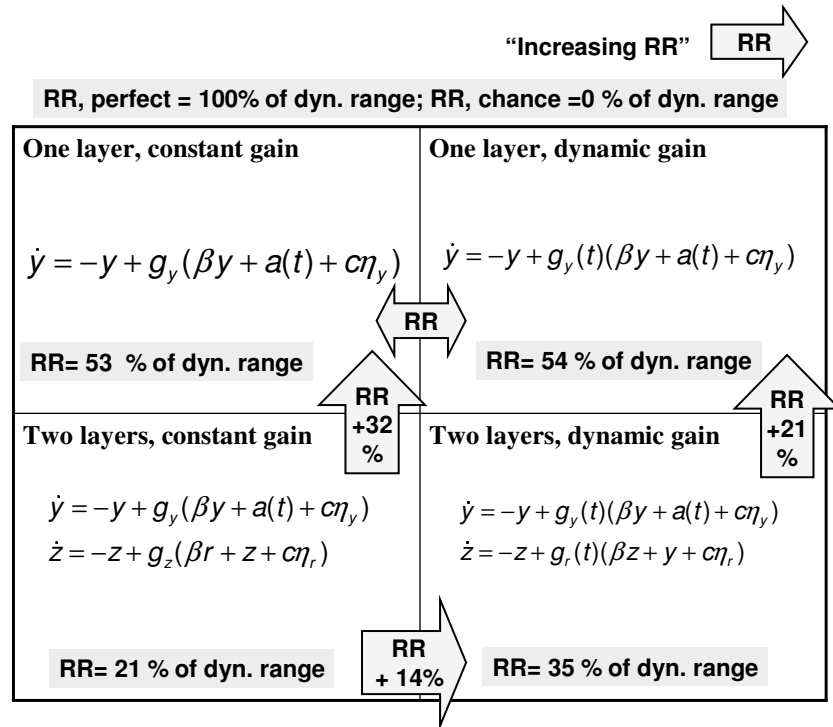
Figure 7.13: Summary of results for the reward rate (RR) optimization problems with motor threshold h allowed to vary freely. For the two-layer models, the optimization problem with adaptive, dynamic gain (7.16) gives higher reward rate (bottom right) than its constant gain counterpart (7.17) (bottom left). However, the optimization problems for the one-layer model yield similar reward rates for both the constant gain ((7.26), top left) and variable gain ((7.25), top right) versions. Reward rates are presented as percentages within the RR 'dynamic range' $[RR_{floor}, RR_{ceil}]$ (Sect. 7.4.3). All results are for the standard parameter set of Sect. 7.4.2.

This strategy of allowing integrated sensory inputs to trigger responses directly is not available for two-layer decision models, in which a second level of (noisy) integration to threshold precedes responses. In this two-layer case, the firing rate trajectories of Figure 7.9 illustrate how transients in gain can streamline the transmission of sensory evidence through the network. The mechanism exploited here is to increase gain in the second layer to a sufficiently high level so as to cause, via positive 'unstable' feedback, a rapid readout of the evidence that has accumulated in the second layer. This gain increase occurs at a delay t_{NE} following crossing of optimally tuned thresholds $\pm h_g$ in the first layer (note that evidence, typically in favor of the 'correct' alternative, continues to accumulate in *both* layers during this delay period).

Despite their optimal use of transient gain schedules, reward rates for two-layer models can still be significantly worse than those achievable by one-layer models (righthand column of Figure 7.13), which achieve an effective 'short-circuit' from sensory input to response. However, this is only the case if motor thresholds are allowed to vary freely. The righthand column of Figure 7.14 illustrates that, when

154

"Increasing RR" RR

RR, perfect = 100% of dyn. range; RR, chance =0 % of dyn. range

One layer, constant gain

$$\dot{y} = -y + g_y(\beta y + a(t) + c\eta_y)$$

RR +11 %

RR= 27 % of dyn. range

One layer, dynamic gain

$$\dot{y} = -y + g_y(t)(\beta y + a(t) + c\eta_y)$$

RR= 38 % of dyn. range

RR +16 %

Two layers, constant gain

$$\dot{y} = -y + g_y(\beta y + a(t) + c\eta_y)$$
$$\dot{z} = -z + g_z(\beta r + z + c\eta_r)$$

RR= 11 % of dyn. range

RR +24 %

RR +3 %

Two layers, dynamic gain

$$\dot{y} = -y + g_y(t)(\beta y + a(t) + c\eta_y)$$
$$\dot{z} = -z + g_r(t)(\beta z + y + c\eta_r)$$

RR= 35% of dyn. range
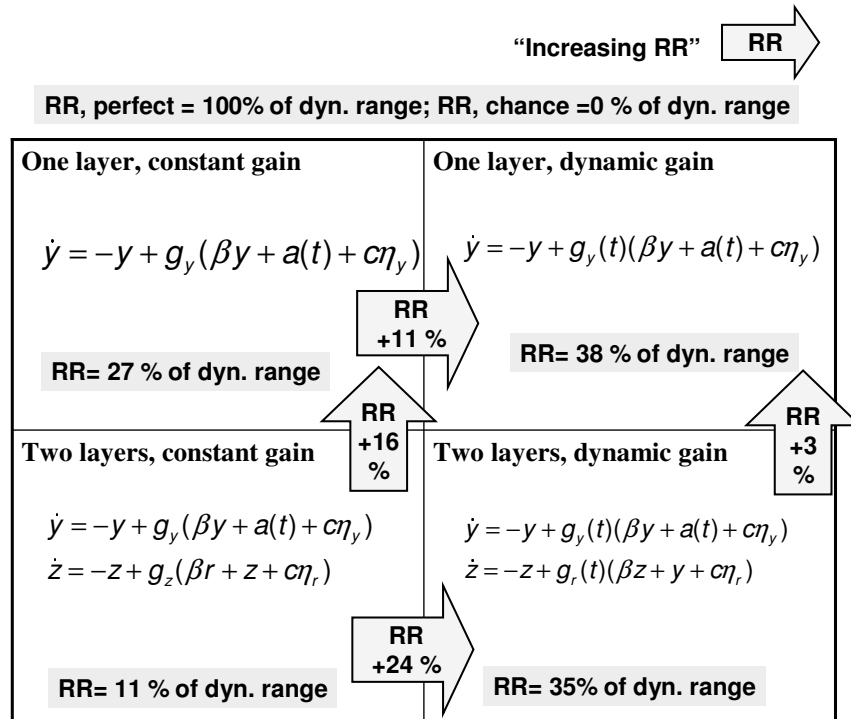
Figure 7.14: As for Figure 7.13, but with fixed motor threshold $h = 5$. In this case, dynamic, adaptive gain provides a significant increase in optimal RR for both one- and two-layer models. Further, the difference in RR for one- and two- layer models with dynamic gain is slight. RR reported as percentages within $[RR_{floor}, RR_{ceil}]$, as in Figure 7.13.

thresholds $\pm h$ for evoking motor responses are fixed, a two-layer model can accomplish a reward rate comparable to that of its one- layer counterpart. In other words, for models with fixed response thresholds, most of the performance apparently lost by adopting a (noisy) two- vs. one- layer architecture can be recovered via LC mediated adaptive gain schedules.

CONCLUSION

Chapter 8

Conclusion

8.1 Summary of dissertation

This dissertation studies mathematical models for cognitive control, or the neural mechanisms by which processing of sensory information is optimized to meet the needs of changing cognitive tasks. The goal is to understand the dynamics of the underlying networks of neurons and the means by which these networks can reproduce and predict features measured in brain recordings and simple behavioral tasks.

First, general results on the individual and population dynamics of neural oscillators are developed. These are applied in later sections, but should be useful in other contexts. In Chapter 2, phase reductions are derived for neural ODEs which show how the four codimension-one bifurcations to periodic firing in single oscillators affect sensitivity to inputs as a function of bifurcation parameters. In Chapter 3, these single-neuron results are extended to study the response of oscillator populations to external stimuli. New results here include general scaling relationships between baseline and stimulated population-averaged frequencies, and the counterintuitive result that certain oscillator populations near bifurcations respond to stimulus offset with greater increases in firing rates than they respond to onset and during stimulus.

In Chapter 4, the dynamics of a finite set of identically (mean field) coupled oscillators are analyzed in general cases of phase-difference coupling and in specific cases of combined phase-difference and phase-dependent 'product' coupling. This chapter adds to existing results based on equivariant dynamics which describe how coupling functions determine the existence and stability of phase-locked states in which subgroups of oscillators are synchronized. New results include domains of attraction for the synchronized solution and new states for systems with a reflection symmetry. Emphasis is placed on the dependence of the stability and degeneracy of phase locked states on the Fourier content of the coupling functions. In particular, while single harmonic sine and cosine functions are degenerate in that they give rise to steady states with multiple zero eigenvalues, the inclusion of higher harmonics generically produces structurally stable equilibria (modulo an overall symmetry) with only a single zero eigenvalue in the relative phase direction. Illustrative examples of coupling functions are derived from Hodgkin-Huxley equations for a single compartment neu-

ron, and their implications for the cross-correlograms often used to assess synchrony in experimental contexts are discussed.

The general results of Chapters 2-4 are applied in Chapter 5, in which a biophysical model for LC neurons is developed and used to explore the role of baseline frequency and stimulus duration on the firing of LC neurons, and hence on neuromodulator release rates. Performing a phase reduction based on the normal form for an individual LC neuron and then taking the continuum (infinite population) limit reveals how timescales inherent to the LC interact with those in the stimuli associated with behavioral tasks. By demonstrating how levels of LC firing in response to stimuli of fixed strength are elevated when the LC has a lower pre-stimulus "baseline" spike rate, these results suggest a new mechanism, supported by recent anatomical discoveries, via which the LC may switch between firing patterns associated with different levels of cognitive performance. The analysis also shows how to "invert" the time course of observed LC firing rates to predict the duration of neural currents representing sensory stimuli in different cognitive tasks. The models of Chapter 5 are compared in detail with the motivating neurophysiological data.

Chapters 6 and 7 move to a more abstracted form of modeling involving neural integrators, in which each variable represents the average activity or firing rate of a pool of neurons selective for a particular stimulus or motor action. I determine how transient parameter adjustments can optimize decision tasks for speed and accuracy in the presence of noise. Chapter 6 shows how either explicit apriori knowledge of the time-course of sensory inputs or information implicit in salient stimulus onset events may be exploited to optimally vary the effective gain on neural units, possibly via transient release of norepinephrine from the LC. In Chapter 7, I investigate how LC mediated gain changes can improve task performance when, by contrast, no knowledge of stimulus timing is available. This requires an adaptive mechanism for driving gain transients based on trial-to-trial information about variable sensory stimuli; this is implemented by increasing gain after a physiologically-imposed delay that follows accumulation of sufficient information to indicate the presence of a stimulus. The results show that, for simple neural integrator networks with two layers representing accumulation of sensory inputs and production of motor responses, LC-mediated adaptive gain changes can significantly improve reward rates achieved in model decision tasks.

8.2 Future directions

This dissertation takes the view that a firing rate description of the averaged 'effective' dynamics of neural subpopulations is sufficient to describe decisions (expressed via motor responses) in certain sensory discrimination tasks (see the Introduction for a summary of some of the relevant experimental data). As also mentioned in the Introduction, interesting questions surround the issue of when additional levels of description are required to capture decision dynamics, and ongoing research (e.g. [135]) here will suggest improvements to decision models for various tasks.

Nevertheless, to the extent that networks whose firing rates integrate sensory evi-

dence over time form adequate building blocks for decision models, we still need a better understanding of how these dynamics emerge from models comprised of large, but finite, numbers of interacting neurons described by sets of nonlinear Hodgkin-Huxley type ODEs having both excitable and oscillatory behavior. The approach required to make this link will significantly extend the probabilistic models of Chapters 3-5, which focussed on the response of uncoupled and weakly coupled neural oscillators to pulsatile inputs. In order to compare the likelihood of alternative hypotheses about sensory inputs, strong inhibitory coupling among subpopulations is required, as for the formal neural integrator models of Chapters 6-7 (cf. [169]). Furthermore, as in the spiking neuron model of [184], a high density of slow excitatory (NMDA) synapses may be required in order for neural populations to integrate noisy inputs over the hundreds-of-millisecond timescales important in sensory discrimination tasks. Analyzing such coupling effects will require extending studies of the existence and stability of partially synchronized firing patterns of Chapter 4 in several ways. The analysis there computed Floquet multipliers characterizing the deterministic multistability of various partially synchronized states. A stochastic stability analysis, performed for Poisson-like fluctuations characteristic of synaptic transmission, would give refined estimates of which among multiple equilibria typically occur in general globally coupled oscillator systems. Moreover, richer coupling architectures and heterogeneous neural populations must also be considered (cf. [63]).

Following these advances, which will lead to population models derived from single-neuron normal forms valid in both excitable and oscillatory regimes (cf. [94]) and containing the synaptic coupling terms necessary to accurately model cortical decision dynamics, there is another, more fundamental, development that will be required in order to bridge between single-neuron and averaged models of simple decisions. This is to develop tractable low dimensional descriptions of the dynamics of coupled normal form populations. A hypothesis which emerges from Chapter 6 and [21, 169, 20] is that the dominant temporal dynamics of neural network equations can be represented as drift-diffusion processes crossing firing rate thresholds on lower dimensional attracting surfaces (slow manifolds). Identifying the corresponding low dimensional, stochastic descriptions of the density equations described above will be a greater challenge. For weakly interacting populations – and hence weakly nonlinear density equations – eigenfunction methods will help to identify a family (likely to be small [165, 104]) of slowly evolving modes that alone determine, to high accuracy, the evolution of the entire population (transients in the other modes decay rapidly). For fully nonlinear and nonequilibrium systems, one could determine the effective dynamics numerically via multiscale "coarse computation" techniques [66], in which the dynamics of the *full* system are regularly sampled and projected onto a proposed set of slow modes. The result will be a reduction of the population equations to a finite set of stochastic ODEs or maps with numerically defined coefficients. These could in turn be used to study the dynamics of decision tasks.

Once low dimensional decision models more closely connected to the underlying neurobiology have been established, the questions about optimality addressed in Chapters 6-7 can be addressed in greater detail. Two issues are of primary importance here. The first is the robustness of optimal decision strategies to imprecise

tuning of parameters. The importance of robustness has long been recognized in the closely related context of neural integrators in motor control (see [155] and references therein), and work in progress by Carlos Brody [19] extends the mechanism of [109] to robustly produce 'line attractors' (of the type required for the optimal decision processing strategies of Chapter 6) in single populations to pairs of mutually inhibitory populations. A logical next step is assessing the extent to which suitable modifications of the networks of [109, 19] can produce robust threshold crossing behavior in timed decision tasks. The second issue directly relevant to this dissertation is that of refining the population level description of norepinephrine (NE) effects. In Chapters 6-7, LC-mediated NE was modelled as causing a simple enhancement of gain in neural populations. While this interpretation is consistent with physiological data, it is not uniquely so. What is required to form an improved model of NE effects on decisions are experiments involving simultaneous recordings from LC and sensorimotor areas (i.e., LIP and superior colliculus), as well as modelling efforts involving populations of spiking cortical neurons equipped with different densities of NE receptors (perhaps following [124]).

Studies of NE and other neuromodulator effects on neural dynamics at this level of detail will also allow an assessment of how neuromodulators influence transmission of information coded via spike times instead of via population or time-averaged firing rates. I conclude this dissertation by noting that, due to the widespread projection of LC neurons to most other brain areas, this fascinating question is almost sure to be relevant to neural computation, and may play a role in unravelling the role of neuromodulator deficits in psychiatric disease.

Bibliography

[1] V. Alvarez, C. Chow, E. van Bockstaele, and J. Williams. Frequency-dependent synchrony in locus coeruleus: role of electronic coupling. *Proc. Nat. Acad. Sci. USA*, 99:4032–4036, 2002.

[2] D. Amit and M. Tsodyks. Quantitative study of attractor neural network retrieving at low spike rates: I. Substrate–spikes, rates, and neuronal gain. *Network*, 2:259–273, 1991.

[3] J. Anderson. *The adaptive character of thought*. Lawrence Erlbaum, Hillsdale, NJ, 1990.

[4] V.I. Arnold. *Ordinary Differential Equations*. MIT Press, Boston, 1973.

[5] D.G. Aronson, M. Golubitsky, and M. Krupa. Coupled arays of Josephson junctions and bifurcation of maps with S_N symmetry. *Nonlinearity*, 4:861–902, 1991.

[6] P. Ashwin and J. Swift. The dynamics of n weakly coupled identical oscillators. *J. Nonlin. Sci.*, 2:69–108, 1992.

[7] G. Aston-Jones, H. Akaoka, P. Charlety, and G. Chouvet. Serotonin selectively attenuates glutamate-evoked activation of locus coeruleus neurons in vivo. *J. Neurosci.*, 11:760–769, 1991.

[8] G. Aston-Jones, S. Chen, Y. Zhu, and M.L. Oshinsky. A neural circuit for circadian regulation of arousal. *Nature Neurosci.*, 4:732–738, 2001.

[9] G. Aston-Jones, J. Rajkowski, and J. Cohen. Locus coeruleus and regulation of behavioral flexibility and attention. *Prog. Brain Res.*, 126:165–182, 2000.

[10] G. Aston-Jones, J. Rajkowski, P. Kubiak, and T. Alexinsky. Locus coeruleus neurons in the monkey are selectively activated by attended stimuli in a vigilance task. *J. Neurosci.*, 14:4467–4480, 1994.

[11] G. Aston-Jones, Y. Zhu, and P. Card. Gabaergic afferents to locus coeruleus (LC) from the peri-LC region: possible LC interneurons. *Soc. Neurosci. Abst.*, 27: 373.8, 2001.

[12] W. Bialek, F. Rieke, R. de Reuter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1854–1857, 1991.

[13] R. Bogacz, E. Brown, J. Moehlis, P. Hu, P. Holmes, and J.D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two alternative forced choice tasks. *In preparation for submission to Psych. Rev.*, 2004.

[14] R. Bogacz, N. Yeung, and C. Holroyd. Detection of phase resetting in the electroencephalogram: an evaluation of methods. Abstract, Society for Neuroscience, Washington DC, 2002.

[15] M. Botvinick, T. Braver, D. Barch, C. Carter, and J.D. Cohen. Conflict monitoring and cognitive control. *Psych. Rev.*, 108:625–652, 2001.

[16] P. Bressloff and S. Coombes. Dynamics of strongly coupled spiking neurons. *Neural Comp.*, 12:91–129, 2000.

[17] P.C. Bressloff and S. Coombes. Spike train dynamics underlying pattern formation in integrate-and-fire oscillator networks. *Phys. Rev. Lett.*, 81:2384–2387, 1998.

[18] K.H. Britten, M.N. Shadlen, W.T. Newsome, and J.A. Movshon. Responses of neurons in macaque MT to stochastic motion signals. *Visual Neurosci.*, 10:1157–1169, 1993.

[19] C. Brody. Personal communication. 2004.

[20] E. Brown, J. Gao, P. Holmes, R. Bogacz, M. Gilzenrat, and J Cohen. Simple networks that optimize decisions. *Int. J. Bifurcation and Chaos*, to appear, 2004.

[21] E. Brown and P. Holmes. Modeling a simple choice task: stochastic dynamics of mutually inhibitory neural groups. *Stochastics and Dynamics*, 1(2):159–191, 2001.

[22] E. Brown, P. Holmes, and J. Moehlis. Globally coupled oscillator networks. In E. Kaplan, J.E. Marsden, and K.R. Sreenivasan, editors, *Problems and Perspectives in Nonlinear Science: A celebratory volume in honor of Lawrence Sirovich*, pages 183–215. Springer, New York, 2003.

[23] E. Brown, J. Moehlis, and P. Holmes. On the phase reduction and response dynamics of neural oscillator populations. *Neural Comp.*, 16:673–715, 2004.

[24] E. Brown, J. Moehlis, P. Holmes, E. Clayton, J. Rajkowski, and G. Aston-Jones. The influence of spike rate and stimulus duration on noradrenergic neurons. *J. Comp. Neurosci.*, 17(1):5–21, 2004.

[25] E.T. Brown, M.S. Gilzenrat, and J.D. Cohen. Adaptive gain and the locus coeruleus. *In preparation*, 2004.

[26] N. Brunel, F. Chance, N. Fourcaud, and L.F. Abbott. Effects of synaptic noise and filtering on the frequency response of spiking neurons. *Phys. Rev. Lett.*, 86:2186–2189, 2001.

[27] C. S. Carter, T. S. Braver, D. M. Barch, M. M. Botvinick, D. Noll, and J. D. Cohen. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280:747–749, 1998.

[28] B. J. Casey, J. D. Cohen, K. O'Craven, R. Davidson, W. Irwin, C. A. Nelson, D. C. Noll, X. Hu, M. J. Lowe, B. R. Rosen, C. L. Truwitt, and P. A. Turski. Reproducibility of fMRI results across four institutions using a working memory task. *Neuroimage*, 8:249–261, 1998.

[29] A.R.R. Casti, A. Omurtag, Sornborger A., E. Kaplan, B. Knight, L. Sirovich, and Victor J. A population study of integrate-and-fire-or-burst neurons. *Neural Comp.*, 14:957–986, 2001.

[30] F.S. Chance, L.F. Abbott, and A.D. Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–782, 2002.

[31] R. Cho, L. Nystrom, E. Brown, A. Jones, T. Braver, P. Holmes, and J. D. Cohen. Mechanisms underlying performance dependencies on stimulus history in a two-alternative forced choice task. *Cognitive, Affective, and Behavioral Neuroscience*, 2(4):283–299, 2002.

[32] C. Chow and N. Kopell. Dynamics of spiking neurons with electrotonic coupling. *Neural Comp.*, 12:1643–1678, 2000.

[33] E. Clayton, J. Rajkowski, J.D. Cohen, and G. Aston-Jones. Decision-related activation of monkey locus coeruleus neurons in a forced choice task. In preparation, 2004.

[34] E. Clayton, J. Rajkowski, J.D. Cohen, and G. Aston-Jones. Decision-related activation of monkey locus coeruleus neurons in a forced choice task. 2004. under preparation.

[35] E. Coddington and N. Levinson. *Theory of ordinary differential equations.* Maple, York, PA, 1955.

[36] J. D. Cohen and T. A. Huston. Progress in the use of interactive models for understanding attention and performance. In C. Umilta and M. Moscovitch, editors, *Attention and Performance XV*, pages 453–476. MIT Press, Cambridge, 1994.

[37] J. D. Cohen and D. Servan-Schreiber. Context, cortex and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psych. Review*, 99:45–77, 1992.

[38] J.D. Cohen, K. Dunbar, and J.L. McClelland. On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, 97(3):332–361, 1990.

[39] J.D. Cohen, D. Servan-Schreiber, and J.L. McClelland. A parallel distributed processing approach to automaticity. *American Journal of Psychology*, 105:239–269, 1992.

[40] M. Cohen and S. Grossberg. Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE transactions on systems, man, and cybernetics*, SMC-13:815–826, 1983.

[41] J. Connor, D. Walter, and R. McKown. Neural repetitive firing: modifications of the Hodgkin-Huxley axon suggested by experimental results from crustacean axons. *Biophys. J.*, 18:81–102, 1977.

[42] J.D. Crawford. Scaling and singularities in the entrainment of globally coupled oscillators. *Phys. Rev. Lett.*, 74:4341–4344, 1995.

[43] J.D. Crawford and K.T.R. Davies. Synchronization of globally coupled phase oscillators: singularities and scaling for general couplings. *Physica D*, 125 (1-2):1–46, 1999.

[44] H. Daido. Generic scaling at the onset of macroscopic mutual entrainment in limit cycles with uniform all-to-all coupling. *Phys. Rev. Lett.*, 73(5):760–763, 1994.

[45] W. E. personal communication. 2001.

[46] R. Eckhorn. Neural mechanisms of scene segmentation: recordings from the visual cortex suggest basic circuits for linking field models. *IEEE Trans. Neural Networks*, 10:464–479, 1999.

[47] B. A. Eriksen and C. W. Eriksen. Effects of noise letters upon the identification of target letters in a non-search task. *Perception and Psychophysics*, 16:143–149, 1974.

[48] G.B. Ermentrout. $n : m$ phase locking of weakly coupled oscillators. *J. Math. Biol.*, 12:327–342, 1981.

[49] G.B. Ermentrout. Reduction of conductance-based models with slow synapses to neural nets. *Neural Comp.*, 6:679–695, 1994.

[50] G.B. Ermentrout. Type I membranes, phase resetting curves, and synchrony. *Neural Comp.*, 8:979–1001, 1996.

[51] G.B. Ermentrout. *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*. SIAM, Philadelphia, 2002.

[52] G.B. Ermentrout and N. Kopell. Frequency plateaus in a chain of weakly coupled oscillators, I. *SIAM J. Math. Anal.*, 15:215–237, 1984.

[53] G.B. Ermentrout and N. Kopell. Oscillator death in systems of coupled neural oscillators. *SIAM J. Appl. Math.*, 50:125–146, 1990.

[54] G.B. Ermentrout and N. Kopell. Multiple pulse interactions and averaging in coupled neural oscillators. *J. Math. Biol.*, 29:195–217, 1991.

[55] L. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, 1998.

[56] A. Fairhall, G. Lewen, W. Bialek, and R. de Ruyter van Steveninck. Effciency and ambiguity in an adaptive neural code. *Nature*, 412:787–792, 2001.

[57] Wong A. Wallace M. Farah M., Brunn J. and Carpenter P. Frames of reference for allocating attention to space: Evidence from the neglect syndrome. *Neuropsychologia*, 28:335–347, 1990.

[58] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, 1968.

[59] N. Fenichel. Persistence and smoothness of invariant manifolds for flows. *Ind. Univ. Math. J.*, 21:193–225, 1971.

[60] E. Fetz and B. Gustaffson. Relation between shapes of post-synaptic potentials and changes in firing probability of cat motoneurones. *J. Physiol.*, 341:387–410, 1983.

[61] R. Fitzhugh. Mathematical models of excitation and propagation in nerve. In H.P. Schwan, editor, *Biological Engineering*. McGraw-Hill, New York, 1969.

[62] S.L. Foote, F.E. Bloom, and G. Aston-Jones. Nucleus locus coeruleus: new evidence of anatomical and physiological specificity. *Physiol. Rev.*, 63 (3):844–914, 1983.

[63] N. Fourcaud and N. Brunel. Dynamics of the firing probability of noisy integrate-and-fire neurons. *Neural Comp.*, 14:2057–2110, 2002.

[64] M. Freidlin and A. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, New York, 1998.

[65] C. Gardiner. *Handbook of Stochastic Methods*. Springer, New York, 1985.

[66] C. W. Gear and Y. Kevrekedis. Projective methods for stiff differential equations: problems with gaps in their eigenvalue spectrum. *SIAM J. Sci. Comp.*, 24:1091–1106, 2003.

[67] W. Gerstner. Population dynamics of spiking neurons: fast transients, synchronous states, and locking. *Neural Comp.*, 12:43–89, 2000.

[68] W. Gerstner and W. Kistler. *Spiking Neuron Models*. Cambridge University Press, Cambridge, 2002.

[69] W. Gerstner, L. van Hemmen, and J. Cowan. What matters in neuronal locking? *Neural Comp.*, 8:1653–1676, 1996.

[70] M.S. Gilzenrat, B.D. Holmes, J. Rajkowski, G. Aston-Jones, and J.D. Cohen. Simplified dynamics in a model of noradrenergic modulation of cognitive performance. *Neural Networks*, 15:647–663, 2002.

[71] L. Glass and M. Mackey. *From Clocks to Chaos*. Princeton Paperbacks, Princeton, NJ, 1988.

[72] J. Gold and M. Shadlen. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.

[73] J.I. Gold and M.N Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Science*, 5 (1):10–16, 2001.

[74] Josh Gold. Personal communication. 2004.

[75] D. Golomb, D. Hansel, B. Shraiman, and H. Sompolinsky. Clustering in globally coupled phase oscillators. *Phys. Rev. A*, 45(6):3516–3530, 1992.

[76] M. Golubitsky and D.G. Schaeffer. *Singularities and Groups in Bifurcation Theory, Vol. 1*. Springer, New York, 1985.

[77] M. Golubitsky, I. Stewart, and D.G. Schaeffer. *Singularities and Groups in Bifurcation Theory, Vol. 2*. Springer, New York, 1988.

[78] S.J. Grant, G. Aston-Jones, and D.E. Redmond. Responses of primate locus coeruleus neurons to simple and complex sensory stimuli. *Brain Res. Bull.*, 21 (3):401–410, 1988.

[79] C.M. Gray. The temporal correlation hypothesis of visual feature integration: Still alive and well. *Neuron*, 24:31–47, 2000.

[80] S. Grossberg. Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.

[81] J. Guckenheimer. Isochrons and phaseless sets. *J. Math. Biol.*, 1:259–273, 1975.

[82] J. Guckenheimer and P.J. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.

[83] D.P. Hanes and J.D. Schall. Neural control of voluntary movement initiation. *Science*, 274:427–30, 1996.

[84] D. Hansel, G. Mato, and C. Meunier. Phase dynamics for weakly coupled Hodgkin-Huxley neurons. *Europhys. Lett.*, 25(5):367–372, 1993.

[85] D. Hansel, G. Mato, and C. Meunier. Synchrony in excitatory neural networks. *Neural Comp.*, 7:307–337, 1995.

[86] A. Herrmann and W. Gerstner. Noise and the PSTH responne to current transients: I. General theory and application to the integrate-and-fire neuron. *J. Comp. Neurosci.*, 11:135–151, 2001.

[87] J. Hertz, A. Krough, and R. Palmer. *Introduction to the Theory of Neural Computation.* Perseus Book Group, New York, 1991.

[88] M. Hirsch, C. Pugh, and M. Shub. *Invariant Manifolds, Lecture Notes in Mathematics v. 583.* Springer, New York, 1977.

[89] A. Hodgkin and A. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117:500–544, 1952.

[90] R. Honeycutt. Stochastic Runge-Kutta Algorithms I. White noise. *Phys. Rev. A*, 45:600–603, 1992.

[91] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558, 1982.

[92] J.J. Hopfield and C.D. Brody. What is a moment? transient synchrony as a collective mechanism for spatiotemporal integration. *Proc. Natl. Acad. Sci. USA*, 98:12821287, 2001.

[93] J.J. Hopfield and C.D. Brody. Simple networks for spike-timing-based computation, with application to olfactory processing. *Neuron*, 37:843852, 2003.

[94] F.C. Hoppensteadt and E.M. Izhikevich. *Weakly Connected Neural Networks.* Springer-Verlag, New York, 1997.

[95] A. Huk, J. Palmer, and M. Shadlen. Temporal integration of motion energy underlies perceptual decisions and response times. *Soc. Neurosci. Abstract*, 2002.

[96] E.M. Izhikevich. Neural excitability, spiking and bursting. *Int. J. Bif. Chaos*, 10:1171–1266, 2000.

[97] E.M. Izhikevich. Phase equations for relaxation oscillators. *SIAM J. Appl. Math.*, 60:1789–1804, 2000.

[98] A. Jadbabaie, N. Motee, and M. Barahona. On the stability of the Kuramoto model of coupled nonlinear oscillators. *Submitted to ACC 2004.*

[99] E. Jodo and G. Aston-Jones. Activation of locus coeruleus by prefrontal cortex is mediated by excitatory amino acid inputs. *Brain Res.*, 768: 327-332, 1997.

[100] E. Jodo, C. Chiang, and G. Aston-Jones. Potent excitatory influence of prefrontal cortex activity on noradrenergic locus coeruleus neurons. *Neuroscience*, 83: 63-80, 1998.

[101] D. Johnston and S. Wu. *Foundations of Cellular Neurophysiology.* MIT Press, Cambridge, MA, 1997.

[102] J. Keener and J. Sneyd. *Mathematical Physiology.* Springer, New York, 1998.

[103] S. Kim and S. Lee. Phase dynamics in the biological neural networks. *Physica D*, 288:380–396, 2000.

[104] B. Knight, A. Omurtag, and L. Sirovich. The approach of a neural population to a new equilibrium: an exact theoretical result. *Neural Comp.*, 12:1045–1055, 2000.

[105] S. Konishi, K. Nakajima, I. Uchida, M. Kameyama, K. Nakahara, K. Sekihara, and Y. Miyashita. Transient activation of inferior prefrontal cortex during cognitive set shifting. *Nature Neuroscience*, 1:80–84, 1998.

[106] N. Kopell and G.B. Ermentrout. Phase transitions and other phenomena in chains of coupled oscillators. *SIAM J. on Appl. Math.*, 50:1014–1052, 1990.

[107] N. Kopell and G.B. Ermentrout. Inhibition-produced patterning in chains of coupled nonlinear oscillators. *SIAM J. Appl. Math.*, 54:478–507, 1994.

[108] N. Kopell, G.B. Ermentrout, and T.L. Williams. On chains of oscillators forced at one end. *SIAM J. on Appl. Math.*, 51:1397–1417, 1991.

[109] A. Koulakov, S. Raghavachari, A. Kepecs, and J. Lisman. Model for a robust neural integrator. *Nature Neurosci.*, 5:775 – 782, 2002.

[110] Y. Kuramoto. *Chemical Oscillations, Waves, and Turbulence.* Springer, Berlin, 1984.

[111] Y. Kuramoto. Phase- and center-manifold reductions for large populations of coupled oscillators with application to non-locally coupled systems. *Int. J. Bif. Chaos*, 7:789–805, 1997.

[112] Y. Kuznetsov. *Elements of Applied Bifurcation Theory, Second Edition.* Springer, New York, 1998.

[113] D.R.J. Laming. *Information Theory of Choice-Reaction Times.* Academic Press, New York, 1968.

[114] E.L. Lehmann. *Testing Statistical Hypotheses.* John Wiley & Sons, New York, 1959.

[115] T. Lewis and J. Rinzel. Dynamics of spiking neurons connected by both inhibitory and electrical coupling. *J. Comp. Neurosci.*, 14:283–309, 2003.

[116] Z. Mainen and T. Sejnowski. Reliability of spike timing in neocortical neurons. *Science*, 268:1503–1506, 1995.

[117] S. Makeig, M. Westerfield, T-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295:690–694, 2002.

[118] I.G. Malkin. Methods of Poincaré and Liapunov in the theory of nonlinear oscillations. *Gostexizdat, Moscow*, 1949.

[119] M.S. Matell and W.H. Meck. Neurophysiological mechanisms of interval timing behavior. *BioEssays*, 22:94–103, 2000.

[120] J.L. McClelland. On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, 86:287–330, 1979.

[121] W.S. McCulloch and W. Pitts. Logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5:115–133, 1943.

[122] Platt. M.L. and P.W. Glimcher. Neural correlates of decision variable in parietal cortex. *Nature*, 400:233–238, 2001.

[123] R.Y. Moore and F.E. Bloom. Central catecholamine neuron systems: Anatomy and physiology of the norepinephrine and epinephrine systems. *Ann. Rev. Neurosci.*, 2:113–168, 1979.

[124] K. Moxon. Work in progress. 2004.

[125] M. Mozer. A connectionist model of selective attention in visual perception. In *Proceedings of the tenth annual conference of the Cognitive Science Society*, pages 195–201. Erlbaum, Hillsdale, NJ, 1988.

[126] J.D. Murray. *Mathematical Biology, 3rd. Ed.* Springer, New York, 2001.

[127] J.D. Murray. *Mathematical Biology I: An Introduction.* Springer, New York, 2002. Third edition.

[128] S. Nichols and K. Wiesenfeld. Ubiquitous neutral stability of splay states. *Phys. Rev. A*, 45(12):8430–8435, 1992.

[129] M.A.L. Nicolelis, D.F. Dimitrov, J. Carmena, R. Crist, G. Lehew, J. Kralik, and S. Wise. Chronic, multi-site, multi-electrode recordings in macaque monkeys. *Proc. Natl. Acad. Sci. USA*, 100:11041–11046, 2003.

[130] D. Nykamp and D. Tranchina. A population density approach that facilitates large-scale modeling of neural networks: analysis and application to orientation tuning. *J. Comp. Neurosci.*, 8:19–50, 2000.

[131] K. Okuda. Variety and generality of clustering in globally coupled oscillators. *Physica D*, 63:424–436, 1993.

[132] A. Omurtag, E. Kaplan, B.W. Knight, and L. Sirovich. A population approach to cortical dynamics with an application to orientation tuning. *Network*, 11:247–260, 2000.

[133] A. Omurtag, B.W. Knight, and L. Sirovich. On the simulation of large populations of neurons. *J. Comp. Neurosci.*, 8:51–63, 2000.

[134] R. O'Reilly and Y. Manukata. *Computational Explorations in Cognitive Neuroscience.* MIT Press, 2000.

[135] L. Osborne, W. Bialek, and S. Lisberger. Time course of information about motion direction in visual area MT of macaque monkeys. *J. Neurosci.*, 24:3210 – 3222, 2004.

[136] A. Papoulis. *Signal Analysis.* McGraw-Hill, New York, 1977.

[137] M. I. Posner and C. R. R. Snyder. Attention and cognitive control. In R. Solso, editor, *Information Processing and Cognition: The Loyola Symposium.* Lawrence Erlbaum, Hillsdale, N.J., 1975.

[138] J. Rajkowski, W. Lu, Y. Zhu, J. Cohen, and G. Aston-Jones. Prominent projections from the anterior cingulate cortex to the locus coeruleus (LC) in rhesus monkey. *Soc. Neurosci. Abst.*, 26:838.15, 2000.

[139] R. Ratcliff. A theory of memory retrieval. *Psych. Rev.*, 85:59–108, 1978.

[140] R. Ratcliff, T. Van Zandt, and G. McKoon. Connectionist and diffusion models of reaction time. *Psych. Rev.*, 106 (2):261–300, 1999.

[141] B. Reddi. Decision making: two stages of neural judgement. *Curr. Biol.*, 11(15):R603–R606, 2001.

[142] J. Rinzel and G.B. Ermentrout. Analysis of neural excitability and oscillations. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling*, pages 251–291. MIT Press, 1998.

[143] J. Rinzel and R.N. Miller. Numerical calculations of stable and unstable periodic solutions to the Hodgkin-Huxley equations. *Math. Biosci.*, 49:27–59, 1980.

[144] J. Ritt. *A Probabilistic Analysis of Forced Oscillators, with Application to Neuronal Response Reliability.* PhD thesis, Boston University, 2003.

[145] J. Roitman and M. Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.*, 22:9475–9489, 2002.

[146] R. Rose and J. Hindmarsh. The assembly of ionic currents in a thalamic neuron I. The three-dimensional model. *Proc. R. Soc. Lond. B*, 237:267–288, 1989.

[147] T. Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, University of Texas at Austin, 1990.

[148] D.E. Rumelhart and J.L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA., 1986.

[149] M. Rush and J. Rinzel. The potassium A-current, low firing rates and rebound excitation in Hodgkin-Huxley models. *Bull. Math. Biol.*, 57:899–929, 1995.

[150] J. Schall. Neural correlates of decision processes: neural and mental chronometry. *Curr. Opin. Neurobiol.*, 2003,.

[151] J. Schall, V. Stuphorn, and J. Brown. Monitoring and control of action by the frontal lobes. *Neuron*, 36:309–322, 2002.

[152] J.D. Schall. Neural basis of deciding, choosing and acting. *Nature Reviews in Neuroscience*, 2:33–42, 2001.

[153] W. Schneider and R. M. Shiffrin. Controlled and automatic human information processing: I. detection, search and attention. *Psych. Review*, 84:1–66, 1977.

[154] D. Servan-Schreiber, H. Printz, and J.D. Cohen. A network model of catecholamine effects: Gain, signal-to-noise ratio, and behavior. *Science*, 249:892–895, 1990.

[155] H.S. Seung, D.D. Lee, B.Y. Reis, and D.W. Tank. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron*, 26:259–271, 2000.

[156] M.N. Shadlen and W.T. Newsome. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiology*, 86:1916–1936, 2001.

[157] M. Shelley and D. McLaughlin. Coarse-grained reduction and analysis of a network model of cortical response. I. drifting grating stimuli. *J. Comp. Neurosci.*, 12:97–122, 2002.

[158] J. Shin, C. Koch, and R. Douglas. Adaptive neural coding dependent on the time varying statistics of the somatic input current. *Neural Comp.*, 11:1083–1913, 1999.

[159] P.L. Smith and R. Ratcliff. Psychology and neurobiology of simple decisions. *Trends in Neurosci.*, 27 (3):161–168, 2004.

[160] R. Stein. A theoretical analysis of neuronal variability. *Biophys. J.*, 5:173–194, 1965.

[161] M. Stone. Models for choice-reaction time. *Psychometrika*, 25:251–260, 1960.

[162] S. Strogatz. From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D*, 143:1–20, 2000.

[163] P. Tass. *Phase Resetting in Medicine and Biology.* Springer, New York, 1999.

[164] D. Taylor and P. Holmes. Simple models for excitable and oscillatory neural networks. *J. Math. Biol.*, 37:419–446, 1998.

[165] A. Treves. Mean-field analysis of neuronal spike dynamics. *Network*, 4:259–284, 1993.

[166] K. Tsang, R. Mirollo, and S.H. Strogatz. Dynamics of a globally coupled oscillator array. *Physica D*, 48:102–112, 1991.

[167] M. Usher, J.D. Cohen, D. Servan-Schreiber, J. Rajkowsky, and G. Aston-Jones. The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283:549–554, 1999.

[168] M. Usher and E.J. Davelaar. Neuromodulation of decision and response selection. *Neural Networks*, 15:635–645, 2002.

[169] M. Usher and J.L. McClelland. On the time course of perceptual choice: The leaky competing accumulator model. *Psych. Rev.*, 108:550–592, 2001.

[170] R.J. Valentino and S.L. Foote. Corticotropin-releasing factor disrupts sensory responses of brain noradrenergic neurons. *Neuroendocrinology*, 45(1):28–36, 1987.

[171] C. van Vreeswijk, L. Abbot, and G.B. Ermentrout. When inhibition not excitation synchronizes neural firing. *J. Comp. Neurosci.*, 1:313–321, 1994.

[172] A. Wald. *Sequential Analysis.* John Wiley & Sons, New York, 1947.

[173] S. Watanabe and S. Strogatz. Constants of the motion of superconducting Josephson arrays. *Physica D*, 74:195–253, 1994.

[174] S. Watanabe and J.W. Swift. Stability of periodic solutions in series arrays of Josephson Junctions with internal capacitance. *J. Nonlin. Sci.*, 7:503–536, 1997.

[175] B. Waterhouse, H. Moises, and D. Woodward. Phasic activation of the locus coeruleus enhances responses of primary sensory cortical neurons to peripheral receptive field stimulation. *Brain Res.*, 790:33–44, 1998.

[176] G.B. Whitham. *Linear and Nonlinear Waves.* Wiley, New York, 1974.

[177] W.A. Wickelgren. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41:67–85, 1977.

[178] K. Wiesenfeld, P. Colet, and S.H. Strogatz. Frequency locking in Josephson arrays: Connection with the Kuramoto model. *Phys. Rev. E*, 57 (2):1563–1569, 1998.

[179] J. Williams, R. North, A. Shefner, S. Nishi, and T. Egan. Membrane properties of rat locus coeruleus neurons. *Neuroscience*, 13:137–156, 1984.

[180] J.T. Williams, D.H. Bobker, and G.C. Harris. Synaptic potentials in locus coeruleus neurons in brain slices. *Prog. Brain Res.*, 88: 167-172, 1991.

[181] H. Wilson and J. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12:1–24, 1972.

[182] A. Winfree. Patterns of phase compromise in biological cycles. *J. Math. Biol.*, 1:73–95, 1974.

[183] A.T. Winfree. *The Geometry of Biological Time.* Springer, New York, 2001. Second Edition.

[184] Wang X-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36:955–968, 2002.

[185] W.Q. Zhu. Stochastic averaging methods in random vibration. *Appl. Mech. Rev.*, 41:189–199, 1988.