

Mathematical Neuroscience
MAT323/APC323 Course Notes
Fall 2012.

Philip Eckhoff and Philip Holmes,
with contributions by
Kolia Sadeghi, Michael Schwemmer and KongFatt Wong-Lin.

Program in Applied and Computational Mathematics,
Princeton University.

December 6, 2012

Contents

1	Introduction: Neuronal properties and the nervous system	<i>(≈ 1 week)</i>	5
1.1	Neuroscience		5
1.2	Neurons		7
1.2.1	Action potentials		8
1.2.2	Synapses		10
1.2.3	Neural coding		11
1.3	The central nervous system		11
1.4	The cerebral cortex		14
1.4.1	Discovery of organization of the cerebral cortex		16
1.4.2	Functional organization of the cerebral cortex		17
1.5	The peripheral nervous system		18
1.6	A note on experimental methods		18
2	Mathematical Tools: ODEs, Numerical Methods and Dynamical Systems	<i>(≈ 2.5 weeks)</i>	20
2.1	Linearization and solving linear systems of ODEs		21
2.2	Liapunov functions		25
2.3	Numerical methods		26
2.3.1	Euler's method		26
2.3.2	The second-order Euler-Heun method		33
2.3.3	The Runge-Kutta method		34
2.3.4	Newton's method for finding zeros		35
2.4	Geometric theory of dynamical systems		36

2.4.1	Stable and unstable manifolds	37
2.4.2	Planar systems of ODEs	39
2.4.3	Center manifolds and local bifurcations	42
2.4.4	Periodic orbits and Poincaré maps	51
2.5	An introduction to neural networks as nonlinear ODEs	54
3	Models of single cells (<i>≈ 2 weeks</i>)	60
3.1	Modeling strategies and scales	60
3.2	Ion channels and membranes	64
3.3	The Hodgkin-Huxley (H-H) equations	69
3.4	Two-dimensional simplifications of HH	74
3.5	Bursting neurons	79
3.6	Propagation of action potentials	84
3.6.1	Traveling waves in a simple PDE	87
3.6.2	Traveling waves in the F-N equation	89
3.7	Concluding remarks	90
4	Synaptic connections and small networks (<i>≈ 2.5 weeks</i>)	90
4.1	Synapses and gap junctions: Towards networks	91
4.1.1	Electrical Synapses	91
4.1.2	Chemical Synapses	92
4.2	Integrate-and-fire models of neurons and synapses	95
4.2.1	Integrate-and-fire models for a single cell	96
4.2.2	Integrate-and-fire models with noisy inputs	99
4.2.3	Implementation of synaptic inputs to IF models	100

4.2.4	A pair of coupled integrate-and-fire neurons	107
4.2.5	A final note on neuronal modeling	111
4.3	Phase reductions and phase oscillator models of neurons	114
4.3.1	Phase response curves	114
4.3.2	Phase models and averaging	118
4.3.3	Phase models and half-center oscillators	119
4.4	Central Pattern Generators	120
4.4.1	A CPG model for lamprey	121
4.4.2	A CPG model for insects	122
5	Probabilistic methods and information theory (≈ 2 weeks)	126
5.1	Stimulus to response maps: Encoding	127
5.1.1	Description of spike trains	127
5.1.2	A primer on probability theory	129
5.1.3	Spike triggered averages	132
5.1.4	A Poisson model for firing statistics	134
5.2	Response to stimulus maps: Decoding	137
5.3	Information theoretic approaches	142
5.3.1	Entropy and missing information	142
5.3.2	Entropy of spike trains and spike counts	144
5.3.3	Relative entropy	149
5.3.4	Mutual information	150
6	Models of decision making (≈ 2 weeks)	152
6.1	Two-alternative forced-choice tasks	152

6.1.1	Neural data from the moving dots task	153
6.1.2	A neural network model for 2AFC: leaky competing accumulators	154
6.2	Derivation of firing rate equations from spiking neuron models	157
6.3	The best way to choose between two noisy signals	164
6.3.1	The sequential probability ratio test	165
6.3.2	Random walks and the continuum limit of SPRT	167
6.4	Introduction to stochastic differential equations	169
6.4.1	Change of variables: Itô's lemma	170
6.4.2	The forward Kolmogorov or Fokker-Planck equation	171
6.4.3	First passage problems and the backward Kolmogorov equation	176
6.4.4	Numerical integration: the Euler-Maruyama method	178
6.5	A return to two-alternative forced-choice tasks	179
6.5.1	The free response protocol	179
6.5.2	The interrogation protocol	180
6.6	Optimal performance on a free response task: analysis and predictions	181

1 Introduction: Neuronal properties and the nervous system (\approx 1 week)

Like its partner, **Classical Mechanics and Chaos**, this course has two main aims: to teach how mathematical models that illuminate an application area can be *constructed*, primarily by describing “classic” examples; and to explain techniques by which these models can be *analyzed*, thus yielding predictions and explanations that can be brought to bear on experimental data. This course deals with models of neurons and neural circuitry, and so we start with a brief review of some key neural phenomena, after noting some courses in which these are described in greater depth.

A basic introduction to neuroscience can be obtained from PSY258/NEU258 Fundamentals of Neuroscience and PSY259ab/NEU 259ab Introduction to Cognitive Neuroscience. Other related courses include NEU501-502, the core course of the Neuroscience Graduate program, which has a laboratory component; MOL408/NEU408/PSY404 Cellular and Systems Neuroscience: fundamental principles at the biophysical, cellular, and system levels, and MOL537/PSY517 Computational Neurobiology and Computing Networks: a course closely related to this one, linked to data from various laboratories in the Molecular Biology and Psychology Departments, but with less mathematical content. PSY330/NEU330 Introduction to Connectionist Models goes far beyond the brief taste that we provide in §6. These courses cover neurobiological and psychological aspects in greater detail than we can, but we will provide more depth and detail on modeling and analytical methods. Finally MOL410/510 Introduction to Biological Dynamics takes a similar approach in modeling to the present course, but covers a wider range of phenomena.

1.1 Neuroscience

In the next 16 pages we collect some background information on neuroscience and the nervous system, with an emphasis on human physiology. Much of the material in this section is taken from Abbot and Dayan [50] and Kandel et al. [123]. **Disclaimer: This sketch does NOT substitute for a proper introductory course in Neuroscience! See the preceding paragraph.**

The nervous system has been studied for millenia, and currently progress is faster than ever, largely due to advances in experimental methods, although theoretical work is playing an increasingly important role. The Greek physician Galen thought that the brain was a gland, from which nerves carry fluid to the extremities. In the late nineteenth century, Golgi and Ramon y Cajal made the first detailed descriptions of nerve cells, and one of Cajal’s drawings is seen in Fig. 1. Harrison soon discovered that the axon and dendrites grow from the cell body in isolated culture. In the mid-nineteenth century, it was discovered that electrical activity in a nerve has predictable effects on neighboring neurons. Pharmacology discovered that drugs affect cells by binding to receptors, which opened the door to discovering the chemical communication at synapses. Neuroscience has now grown into a broad and diverse field: molecular neuroscience studies the detailed structure and dynamics of neurons, synapses (their connections), and small networks; systems neuroscience studies larger-scale networks that perform tasks and interact with other such networks (or brain areas) to form pathways for higher-level functions; and cognitive neuroscience studies the relationship between the underlying physiology and behavior, thought, and cognition.

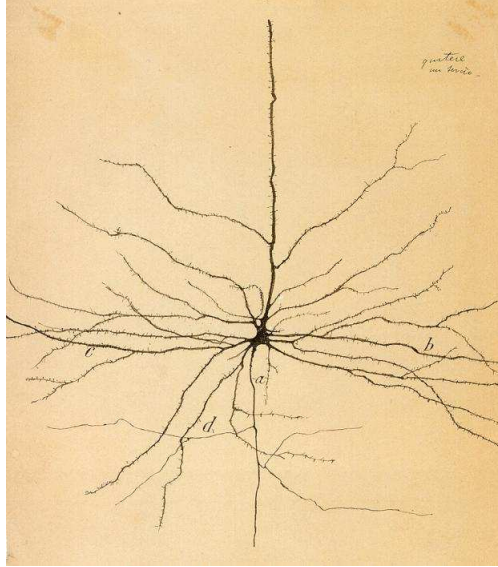


Figure 1: Ramon y Cajal’s sketch of a neuron. (*expasy.org*).

Mathematical treatments of the nervous system began in the mid 20th century. One of the first examples is the book of Norbert Wiener, based on work done with the Mexican physiologist Arturo Rosenblueth, and originally published in 1948 [229]. Wiener introduced ideas from dissipative dynamical systems, symmetry groups, statistical mechanics, time series analysis, information theory, and feedback control. He also discussed the relationship between programmable digital computers (then in their infancy) and neural circuits, a theme that John von Neumann subsequently addressed in a book written in 1955-57 and published in the year after his death [217]. In fact, while developing one of the first programmable digital computers (JONIAc, built at the Institute for Advanced Study in Princeton after the second World War), von Neumann had “tried to imitate some of the known operations of the live brain” [217, see the Preface by Klara von Neumann]. It is also notable that, in developing cybernetics, Wiener drew heavily on von Neumann’s earlier works in analysis, ergodic theory, computation and game theory, as well his own studies of Brownian motion (now known as Wiener processes). We will describe some of these ideas and methods in §§5-6.

The books of Wiener and von Neumann [229, 217] were directed at the brain and nervous system *in toto*, although much of the former was based on detailed experimental studies of heart and leg muscles in animals. The first cellular-level mathematical model of a single neuron was developed in the early 1950’s by the British physiologists Alan Hodgkin and Andrew Huxley [104]. It grew out of a long series of experiments on the giant axon of the squid *Loligo* by Hodgkin, Huxley and many others, as described in §3. Since this early period mathematical neuroscience has grown into a subdiscipline, served worldwide by long and short courses, with a growing list of textbooks (e.g. [231, 50, 124, 61]). The number of published mathematical models must now exceed the catalogue of brain areas by several orders of magnitude. Hopefully, this course will prepare you to profit from, and contribute to, this exciting field.

Before beginning our rapid tour through the physiology of the nervous system, some remarks about mathematical models are necessary. In spite of the growth of modeling, neuroscience is still dominated by experimental methods and findings. Most models are developed to explain and/or to

predict experimental observations. To do so they must be validated by recreating, at least qualitatively, behavior observed experimentally. Models can be of two broad types: *empirical* (also called *descriptive* or *phenomological*), or *mechanistic*. Empirical models ignore the (perhaps unknown) anatomy and physiology, and attempt to reproduce input-output (or stimulus-response) relationships of the biological system under study. Mechanistic models attempt to describe physiological and anatomical features in some detail, and reproduce observed behaviors by appropriate choice of model components and the parameters characterizing them, thereby revealing mechanisms responsible for those behaviors. Many models are not so easily classifiable, and models can also occupy all kinds of places in a continuum from molecular to organismal scales. We shall have more to say about this in §3.1. Most models in neuroscience and mathematical biology strive for quantitative accuracy, but this remains more elusive than is typical in the physical sciences.

1.2 Neurons

The nervous system consists of the peripheral and central nervous systems, which themselves can be further subdivided and studied. The neuron is the building block for the entire system, and so, following the reductive method that has been so successful in the physical sciences, we shall begin with a single cell, and build up to larger systems.



Figure 2: Photograph of neurons from an ox spinal cord smear, with an H & E stain, magnified 100X. (*dmacc.edu*).

We start with some basic anatomy and architecture. Neurons are cells that can carry signals rapidly over distance (albeit at speeds much less than that of light!). These signals are *action potentials* (APs, or spikes): relatively large ($\mathcal{O}(100)$ mV)¹ fluctuations in the voltage across the cell membrane. APs propagate down the axon: an extension of the cell that takes the form of a relatively long cable, compared to the size of the soma, or central cell body. Neurons exhibit significant variety in shape and size, but all have the same basic features: the soma, dendrites,

¹The notation $\mathcal{O}(n)$ means, of the order n , a rigorous definition is provided in §2.1. Here we use it informally to mean “within an order of magnitude.”

which are the shorter and more numerous branching extensions² that receive signals from other neurons, and the axon, which carries signals to other neurons, and may also be branched. Principal (or projection), neurons are often excitatory, and carry signals over long distances from one region of the central nervous system (CNS) to another or from one processing region of the brain to another. Cortical pyramidal neurons, examples of projection neurons, are the primary excitatory neurons of the cerebral cortex, the brain region where most cognitive functions are located. Interneurons have shorter axons and usually provide inhibitory inputs. Fig. 2 shows neurons in a vertebrate spinal cord.

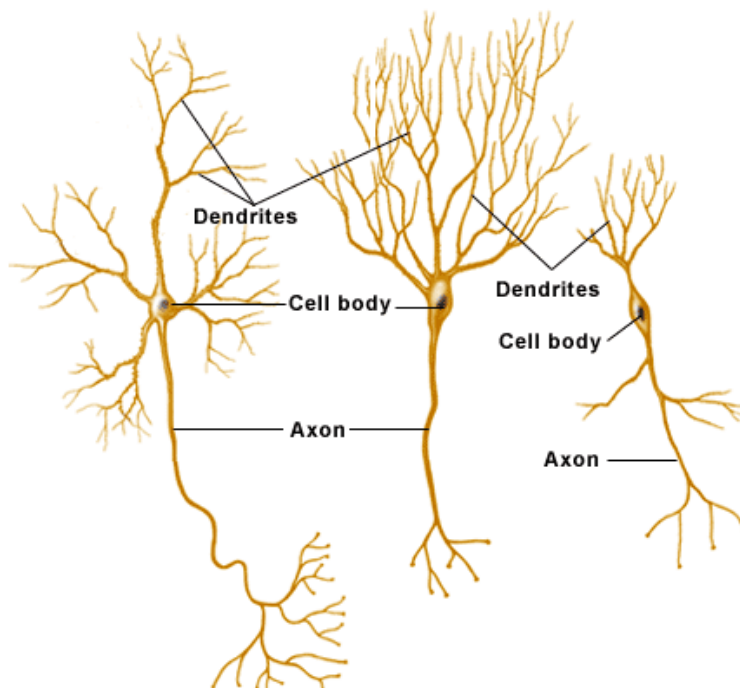


Figure 3: Schematic of parts of a neuron. (college.hmco.com/psychology/bernstein).

The dendritic tree (the branching pattern of dendrites) receives an average of 2 inputs/ μm , and on average each neuron has 4 mm of dendritic tree. The type and number of inputs varies significantly, and some neurons receive over 10^5 inputs. An axon has 180 connections/mm length on average, and a typical mouse neuron has a 40 mm axon. The longest axons in the human can exceed 1 m (extending from motoneurons in the spinal cord to muscles in the feet). Each cell can synapse with thousands of others ($\mathcal{O}(10^3)$) and some brainstem cells make $\mathcal{O}(10^5)$ connections. The human brain contains $\mathcal{O}(10^{11})$ neurons and $\mathcal{O}(10^{14})$ connections. A diagram of the major components of a neuron is given in Fig. 3, and Fig. 4 shows a sample of neuron types, which can vary widely in geometry and types of ionic channels.

1.2.1 Action potentials

The membrane that encloses the neuron contains a large number of ion channels, pores that allow charged ions to pass across it. Most are selective for a specific ion, with Na^+ , K^+ , Ca^{2+} , and

²Biologists refer to dendrites and axons as *processes*: confusing terminology for a mathematician!

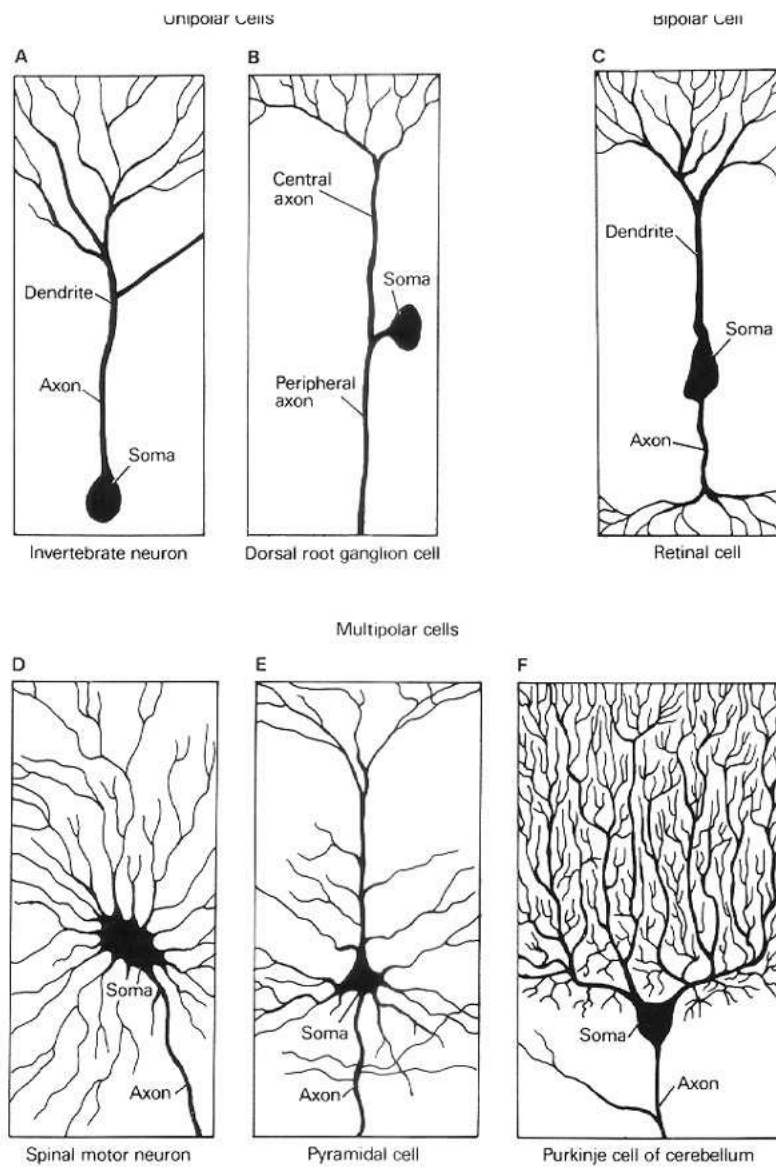


Figure 4: Some different types of neurons. (*students.uni-marburg.de*).

Cl^- the most common. These pores control the flow of each ion by opening and closing in response both to the voltage across the membrane and to internal or external signals. The membrane voltage (or more correctly, trans-membrane voltage) – the difference between the electrical potential inside and outside the cell – is the key variable. A typical *resting potential* (the potential inside the neuron relative to the extracellular fluid measured when the neuron is not spiking) is approximately -70 mV. An *action potential* (AP) or *spike* is a fluctuation on the order of 100 mV that lasts about a millisecond. The resting potential is maintained by ion pumps, which remove Na^+ from the cell and bring K^+ into the cell. Apart from the pumps, ions flow along concentration gradients from high to low (provided their channels are open), and along voltage gradients, negative ions moving towards higher potential, and positive ions towards lower potential. A positive current is defined outward, and consists of positive ions flowing out or negative ions flowing in. A positive current *hyperpolarizes* the cell (makes the voltage more negative), and a negative inward current *depolarizes* it (makes the voltage more positive)³.

Depolarization leads to spikes, and if the depolarization is large enough, the cell will spontaneously spike, with the spike voltage fluctuation often at least an order of magnitude larger than the original depolarization. The spike is typically caused by Na^+ channels opening quickly and allowing an inrush of Na^+ ions, which raises the voltage and creates the spike, followed by K^+ channels opening, allowing an outflow K^+ , while the Na^+ channels close, which drops the membrane voltage below resting potential. While the pumps work to reestablish concentration gradients which enable the spike, there is a short period of time, the *absolute refractory period*, during which the cell cannot spike, followed by a *relative refractory period* as ion concentrations recover, during which it is very difficult, but not impossible, to spike.

Small changes in voltage insufficient to trigger a spike are called *subthreshold potentials*. These are strongly attenuated along the axon and usually cannot be detected even 1 mm away from the soma. Action potentials, on the other hand, regenerate along the axon and travel long distances without attenuation. Wave propagation in this active medium, accompanied by waves of ionic transport across the cell membrane, is quite different from propagation through a passive medium like that of sound waves in still air and electrical pulses in wires. Some cells (e.g., bipolar cells in the retina) typically exhibit only non-spiking subthreshold oscillations.

1.2.2 Synapses

Neurons communicate via *synapses* (sometimes called chemical synapses) and direct electrical contacts called *gap junctions*. Gap junctions influence the voltages of neighboring cells much as if they were connected by a simple electrical resistor.

Synapses are points at which an axon of the presynaptic cell is in close contact with a dendrite of the postsynaptic cell. When an AP arrives at a synapse, channels in the presynaptic terminal open to allow an influx of Ca^{2+} from the extracellular medium. The increased internal calcium concentration causes vesicles that contain neurotransmitter molecules to fuse with the cell membrane, thereby releasing neurotransmitter molecules that diffuse across the synaptic cleft. The other (postsynaptic) side of the synapse is a dendritic spine (a projection from a dendrite). The neuro-

³More confusing terminology.

transmitters open ion channels on the spine, thereby causing currents that either excite (depolarize) or inhibit (typically hyperpolarize) the post-synaptic neuron. A given synapse is either excitatory or inhibitory and the resulting signals are called *excitatory* (resp. *inhibitory*) *post-synaptic potentials*, or EPSPs (resp. IPSPs) for short.

1.2.3 Neural coding

Neural coding is a major field in itself, and in §5 we will spend some time learning the mathematics used and the questions asked. The problem has two components: *encoding* and *decoding*.

Close your eyes and run a finger along a rough surface. What happens as you ‘feel’ the texture and try to identify the material? Force and displacement sensors responding to continuously varying signals (stimuli) cause neurons in the peripheral nervous system to spike, conveying discrete sequences of APs to the somatosensory cortex (see §§1.4-1.5). From these sequences your brain ‘reconstructs’ the surface. In parallel with this process, encoding tries to understand the production of spike sequences produced by a neuron in response to a stimulus. Remember that individual spikes are separated by refractory periods, and the stimulus can be changing rapidly, perhaps on the order of the interspike interval. Decoding attempts to recover the stimulus from a spike sequence. Given the variability in neurons and the sparsity of spikes compared to changes in the stimulus, encoding and decoding are often studied in terms of *firing rates*, which translate into the probability of a spike occurring in a small time interval for an individual neuron, or into the average firing rate at a given time for a population of neurons receiving a common stimulus. However, there is increasing emphasis on the role of individual spike timing [179].

1.3 The central nervous system

The central nervous system (CNS) is bilateral and has seven major divisions: the spinal cord, the medulla oblongata, pons and midbrain (which form the brain stem), the cerebellum, diencephalon, and the cerebral hemispheres: see Fig. 5.

The anatomy of the CNS and its divisions is referenced by axes with directions defined as rostral, caudal, ventral, and dorsal. In most animals, the axes directions are consistent for each division and for the CNS as a whole. Rostral means towards the front (head), caudal towards the tail, ventral towards the belly, and dorsal towards the back. In a rat, for example, these axes maintain orientation for the CNS as a whole, and for the cerebrum locally, as seen in Fig. 6. In higher primates, there is an orientation change above the spinal cord, as seen in Fig. 7. For the CNS as a whole and the spinal cord, rostral is towards the head, caudal towards the coccyx, ventral towards the belly and dorsal the back. Above the spinal cord, and in the brain, rostral is rotated and is towards the front, caudal towards the back of the head. Dorsal is towards the top and ventral towards the bottom, although these two terms are often replaced by superior and inferior. For the remainder of the course, we will try to use these four terms to indicate location and direction, but their orientation depends on the system being discussed, and we may be inconsistent anyway, so stay alert.

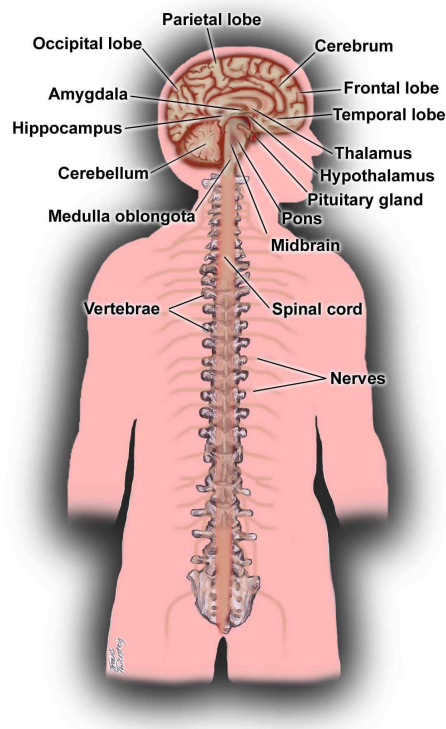


Figure 5: The central nervous system. (*emedicinehealth.com*).

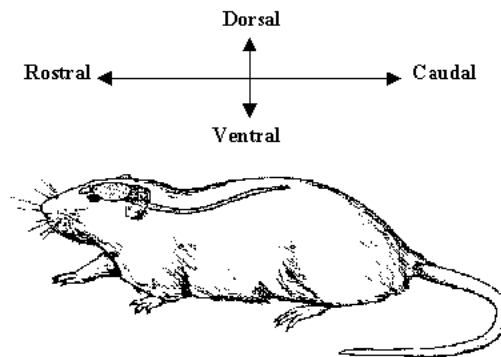


Figure 6: Central nervous system orientation and axes for a rat, typical of most vertebrates. (*fmrrib.ox.ac.uk*).

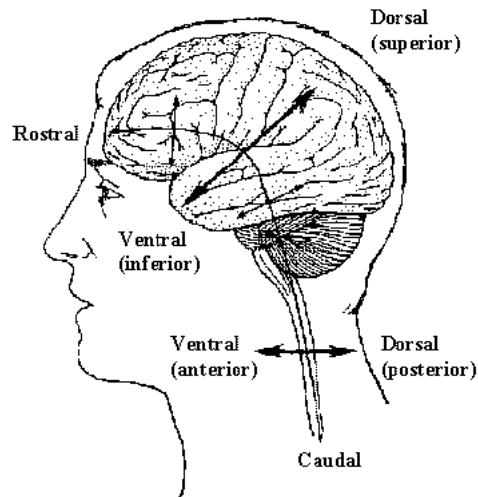


Figure 7: Central nervous system orientation and axes for humans, typical of higher primates. (*fmri.ox.ac.uk*).

Spinal cord: The spinal cord synapses with nerves leading into the limbs and the trunk, allowing control of their movement and reception of sensory information. A cross-section of spinal cord reveals an ‘H’ of gray matter surrounded by white matter, as seen in Fig. 8. The ventral horn of the gray matter (the leg of the ‘H’ pointing towards the belly) contains the cell bodies of motor neurons that control muscles. The dorsal horn contains sensory neurons that receive inputs from the trunk and limbs. The white matter is mostly axons that run up and down the spinal cord, the white actually being the myelin covering the axons.

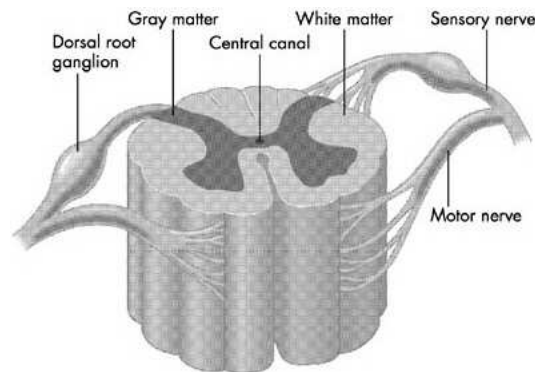


Figure 8: Cross section of a spinal cord, revealing gray matter and white matter. (*nap.edu*).

Brain stem: The brain stem lies at the rostral end of the spinal cord and consists of the *medulla oblongata*, the *pons*, and the *midbrain*. As a whole, it controls the muscles of the face and head, receives sensory signals from the face and head, conveys information between the brain and the spinal cord, and regulates levels of arousal and awareness. The medulla oblongata governs vital autonomic functions such as digestion, breathing, and heart rate. The pons is the link between the cerebrum and the cerebellum. Finally, the midbrain contains sensory and motor function of the head and face, controls eye movement, and is part of the pathway for visual and auditory reflexes.

Cerebellum: The cerebellum controls the force and range of movements and is important in

learning motor skills. The cerebellum actually contains more neurons than any other division of the CNS, even the cerebrum.

Diencephalon: The diencephalon contains the *thalamus* and *hypothalamus*. The thalamus passes sensory information to the cerebral cortex from the rest of the CNS, but it is more than a relay. It has a gating and modulatory function, governing which sensory inputs receive conscious attention. The hypothalamus regulates autonomic, endocrine, and visceral function, as well as many other functions through its control of the pituitary gland. The hypothalamus is also related to motivation and pursuing actions which are rewarding. The hypothalamus can cause the sensation of hunger, for example, and is involved in the complicated process of arousal and selective attention. Dopaminergic neurons in the midbrain are involved in perception of reward, and are also part of the larger attention process. In this way, these regions exert strong modulatory effects on the entire CNS and body.

Cerebral hemispheres: The two cerebral hemispheres contain the *cerebral cortex*, which on each side consists of four lobes (frontal, parietal, temporal, and occipital), the *basal ganglia*, *hippocampus*, and *amygdaloid nuclei* (a nucleus refers to a cluster of neurons). The basal ganglia regulate movement and are involved in cognitive processes such as learning skills. The hippocampus relates to memory, although it is not a memory storage organ. It is responsible for *formation* of long-term memories, and damage does not affect existing memories, but degrades the ability to make new ones. The amygdaloid nuclei, or amygdala, governs autonomic and endocrine responses to emotional states. It discerns any emotional and motivational component to sensory stimuli, and its projection to the brain stem allows it to affect the body. For instance, when experiencing fear, the amygdala is part of the pathway that results in increased heart rates. The cerebral cortex is very complicated and will be treated in its own section.

1.4 The cerebral cortex

As in §1.2 we start with some anatomy. The cerebral cortex is the furrowed ‘gray matter’ which allows cognitive function. It also has a bilateral structure and each hemisphere is divided into four lobes. The *frontal lobe* is an essential part of the pathway for motor control, and it is concerned with future actions. The *parietal lobe* processes somatic sensation, which also involves understanding the current position and orientation of the body and its parts and the relation of the body to space. The *occipital lobe* is mostly occupied with processing vision. The *temporal lobe* processes auditory inputs, houses the hippocampus and amygdaloid nuclei, and is part of visual memory. The lobes of the brain are diagrammed in Fig. 9.

The two hemispheres are connected by the *corpus callosum*, a thick bundle of mostly white matter which allows information to pass between hemispheres. The surface of the cortex has many distinctive grooves and ridges, which most people think of when they think of ‘brains.’ The grooves are called *sulci* and the ridges *gyri*. Most cognitive processes reside in the cortex, which is typically only 2 to 4 mm thick. Therefore, the number of neurons available for cognitive processes is proportional to cortical area. The numerous deep infoldings provide a way to increase the cortex area in a given volume. The extent of infolding varies from species to species, increasing in the higher primates and being most evident in humans. The central sulcus lies between the parietal and frontal lobes, dividing sensory and motor areas. The lateral sulcus defines the temporal lobe, separating it from the frontal and parietal lobes. In addition to the four lobes, there are two major

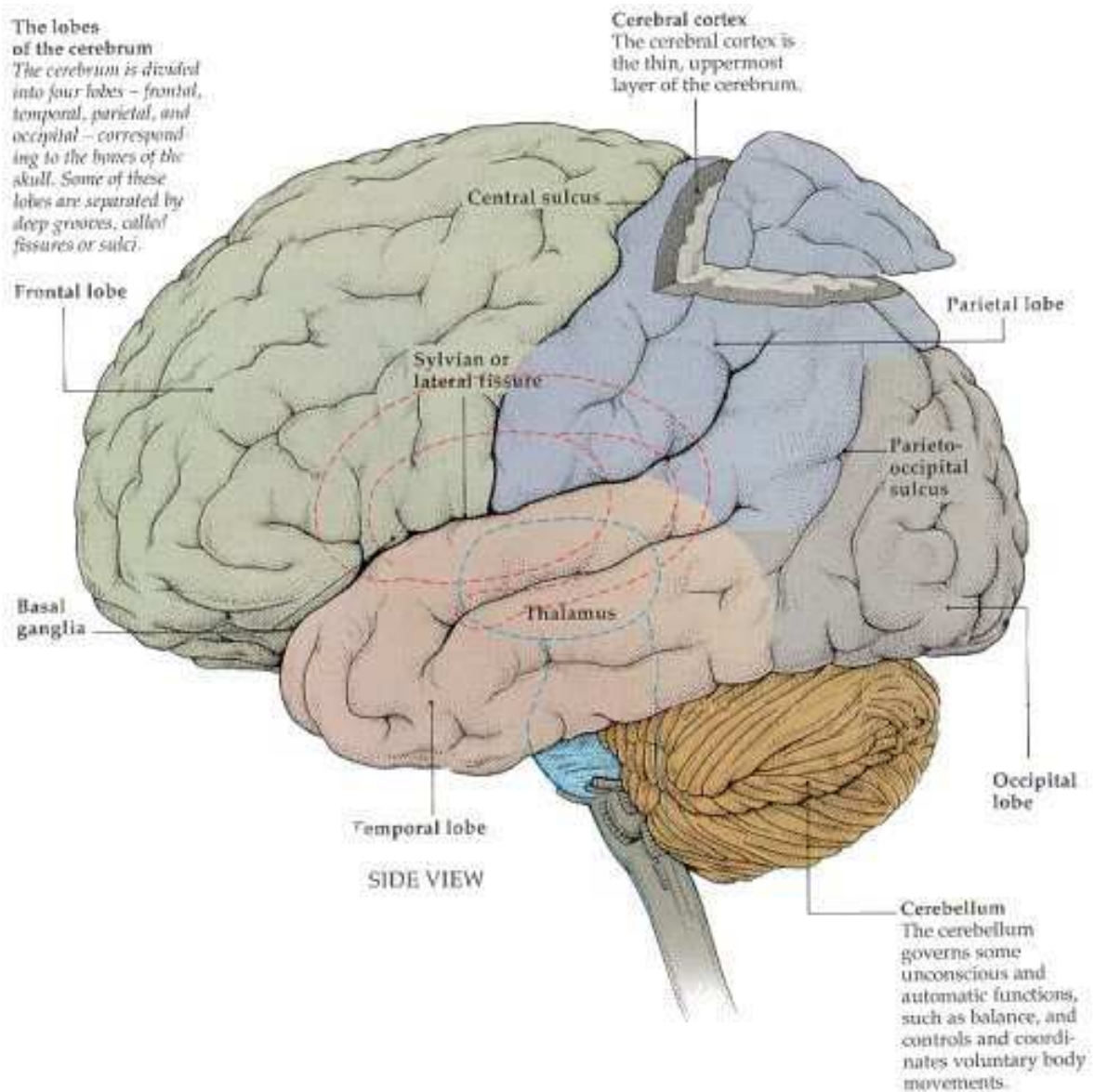


Figure 9: The four lobes of the brain and other prominent structural features. (sruweb.com/-walsh).

cortical areas that are not part of any specific lobe. Above the corpus collosum is the *cingulate cortex*, and inside the lateral sulcus is the *insular cortex*. The insular cortex cannot be seen from outside the brain, since the parietal and temporal lobes bulge around it.

The cerebral cortex exhibits contralateral function: each hemisphere is primarily concerned with the opposite side of the body. The left side of the brain receives input from and controls the right hand. Vision is slightly different, the signals are not distributed depending on what eye they are from, but rather which side of the body they see: the part of each retina that sees the left half-plane sends its signals to the right hemisphere and vice-versa. Recall that in the spinal cord, inputs from the right side go into the right side of the spinal cord, but in the brain they are switched. Between entering the dorsal horn of the spinal cord and arriving at the cerebral cortex, the pathway switches sides. These switches are called decussations and they occur at different points in the ascent, depending on the pathway. Surprisingly, the cerebral hemispheres themselves are not symmetrical: there are deviations from symmetry not only in structure, but also function.

1.4.1 Discovery of organization of the cerebral cortex

The function of the brain has been studied and debated seriously for about two centuries. Gall theorized that different regions of cortex perform different functions, to such an extent that he divided the brain into 35 separate organs, each with a different mental function. He thought that use of certain faculties would increase the size of the corresponding section of cortex, as exercise does for muscles. His ideas led to the pseudoscience of phrenology, in which personal traits and abilities are ascertained from bumps in the skull supposedly caused by changes in these regions of cortex. In the 1820's, Flourens tested this theory experimentally by removing different regions from animals. Removing the entire region that Gall claimed corresponds to a certain mental ability did not destroy that ability in these animals. Flourens decided that the entire brain was involved in each mental function and introduced the aggregate-field view of the brain.

In the mid-nineteenth century, Jackson, and later Wernicke, Ramon y Cajal, and Sherrington showed that sensations and motions in different body parts correspond to different cortical areas. Jackson worked with epilepsy, and saw that convulsions in different parts of the body can be traced to specific and different parts of cortex. The later three developed a new view called cellular connectionism: neurons are basic signalling units, and are arranged in functional groups which connect to other functional groups in an ordered sequence. At the start of the 20th century, Brodmann distinguished 52 functionally distinct areas of cortex with his cytoarchitectonic method. In spite of mounting and compelling evidence for distinct functional areas by the start of the twentieth century, the aggregate-field view dominated the field for several more decades, primarily due to the pronouncements of prominent neural scientists, including Head, Goldstein, Pavlov, and Lashley. Lashley tested for distinct areas by looking for the area that allows rats to find their way through a maze. No matter which region he lesioned, the rats were able to find their way through the maze, and Lashley considered this proof of aggregate-field. In retrospect, we see that many faculties have parallel pathways which can compensate for damage to one of the routes through the brain. For instance, when a rat's vision system was damaged, it was able to find its way through the maze using whiskers. The unscientific nature of phrenology probably played a significant role in the scientific opposition to distinct areas. As the century continued, the evidence was finally sufficiently overwhelming to change the dominant scientific view. For example, a small lesion can

destroy the ability to recognize people by name but not by sight.

Until recently, everything we knew about function of different brain regions was learned in lesion studies, in which physical damage to the brain is observed to correlate with losses in specific functions. For instance, in 1861, Broca was studying aphasia, a language disorder. He had a patient whose mouth, vocal cords, and tongue appeared physically intact, but who could not speak coherently. This man could understand language and utter isolated words and hum music, but he could not speak. Following the patient's death, Broca studied his brain and found a lesion in the left hemisphere. Over the next three years, eight similar cases were found to have lesions in the same spot. This led to Broca's famous announcement that we speak with the left hemisphere. Wernicke later discovered that Broca's region is not the entire language region, but there is another part of the left hemisphere that allows comprehension of language. Later it was discovered that lesions to Broca's area make deaf people unable to sign, even when they know sign language before damage. Today, it is possible to more finely subdivide cortical function, and we can now study distinct regions in fully functional, healthy brains without lesions using imaging methods (see §1.6). For example, the 5 visual processing areas noted by Brodmann have now been subdivided into 35 functionally distinct regions.

In summary, specific brain regions are not concerned with individual mental faculties, as Gall had supposed. Each region performs a basic processing task, but they are linked in both series and parallel, with "reentrant" (two way) connections, to make higher faculties such as language possible. Damage to a single area does not necessarily destroy the faculty, since parallel pathways can compensate, and the brain can even reorganize linkages in some cases.

1.4.2 Functional organization of the cerebral cortex

The cerebral cortex is 2 to 4 mm thick across species, and is arranged in six distinctly-defined *layers*, which are structurally different: some are home to the main cell bodies, some for projection to later regions, some for inputs, some for back-projection to previous regions. The layered structure allows for efficient space utilization and organization of input-output and feedback connections. Basic processing takes place in *cortical columns*, each of which would fit in a 1 mm diameter cylinder. These columns are viewed as the fundamental computational modules, and not surprisingly, humans have many more of them than rats.

The primary *visual cortex* lies in the occipital lobe near the calcarine sulcus. The primary *auditory cortex* is in the temporal lobe on gyri on the lateral sulcus, which separates the temporal lobe from the parietal and frontal lobes. Somatosensory processing is caudal to the central sulcus on the post-central gyrus in the parietal lobe. The primary motor cortex is rostral to the central sulcus, just opposite the somatosensory region, as seen in Fig. 10. The cortex contains topographical representations of the retina, cochlea, and body itself. Thus, that neighboring sections of auditory cortex represent neighboring sections of cochlea, for example. The sensory and motor cortex each have well-defined body maps including arms, hands, even fingers. These maps do not have equal representation, but cortex area is distributed in proportion to the density of sensors or the fineness of movement required in the represented area. For example, the fingers are enormous and completely out of normal proportion in this somatosensory body map. Higher order areas do not exhibit as clearly-defined a map, although topographical representation may still be present. Higher order

motor areas, for example, are located rostral to primary motor cortex in the frontal lobe. Higher order vision areas are rostral to primary visual cortex, area V1. Note that primary somatosensory is the first stop in the sensory pathway into cortex, but the primary motor cortex is the last stop in the motor pathway out of cortex.

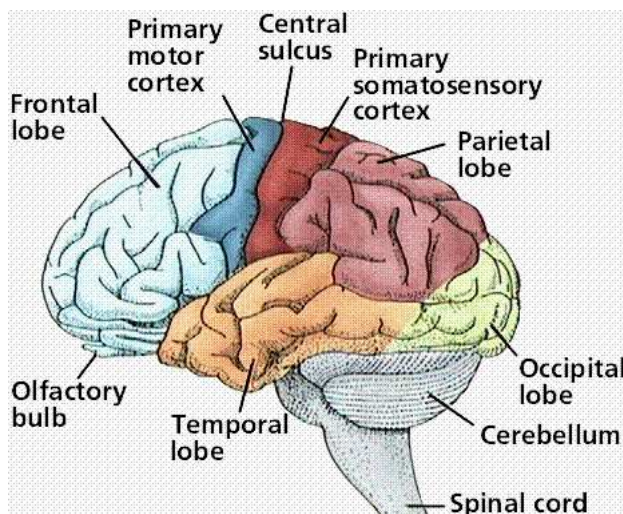


Figure 10: Locations of primary motor and somatosensory cortices, with some other lobes and brain areas indicated. (*emc.maricopa.edu*).

1.5 The peripheral nervous system

The peripheral nervous system has somatic and autonomic functions. Somatic refers to the nerves and nerve endings in skin, muscles, and joints. Sensory information from nerve endings relays to the spinal cord information about the current configuration of the body and about its surroundings. Motor control of muscles is also partially achieved through nerves of the peripheral nervous system. Such feedback, originating in sensors embedded within the body that monitor body states, is called *proprioceptive*.

Autonomic functions include visceral sensation, motor control of viscera, motor control of smooth muscles, and control of exocrine functions. The autonomic functions can be divided into three systems. The *sympathetic system* governs the body's response to stress, the *parasympathetic system* governs conservation of body resources and homeostasis, and the *enteric system* controls the smooth muscle in the gut. Functions of the autonomic nervous system are not consciously controlled. In contrast, *exteroceptive* sensing involves monitoring the external environment via the senses of sight, hearing, touch, taste and smell, which typically requires brain activity, and may rely on conscious reflection.

1.6 A note on experimental methods

Membrane voltage recordings are essential for study of individual neurons and they played a central role in the development of the Hodgkin-Huxley equations (see §3.3). Voltage recordings,

or more general experiments on the neural system, are carried out via two major avenues *in vivo* experiments on live animals, and *in vitro* experiments on slices or aggregates of cells removed from the organism and maintained under suitable conditions.

Changes in membrane potential can be recorded intracellularly or extracellularly. Intracellular recordings are usually done for *in vitro* experiments. Either a sharp electrode pierces the membrane, achieving electrical contact with the intracellular fluid, or the electrode attaches to the outside of the cell, sealing a patch of membrane which is then broken. The electrode typically is attached at the soma, but new techniques allow measurement in the dendritic tree as well. Intracellular recordings are usually not performed on the axon, but when they are, only spikes are discernable, since other voltage fluctuations are attenuated. Extracellular recordings are used for *in vivo* experiments, and do not damage the cell membrane. The disadvantage is that the voltage fluctuations measured during a spike are three orders of magnitude less than for intracellular recordings. Spikes are detected, but not subthreshold fluctuations, and spikes from neighboring cells may also appear on the voltage trace. This has led to sophisticated ‘spike-sorting’ methods and algorithms. Membrane voltages can also be recorded optically, using voltage sensitive dyes.

Neuroanatomical tracing techniques have provided improved techniques for tracing axonal projections, and thereby establishing connectivity among brain areas. Originally, researchers cut an axon and looked for the affected cell bodies. Now, particles can be passed from axon to soma, or from soma to axon, and often from cell to cell. Horseradish peroxidase and the herpes simplex virus are both used, in addition to dyes or radioactively labeled amino acids.

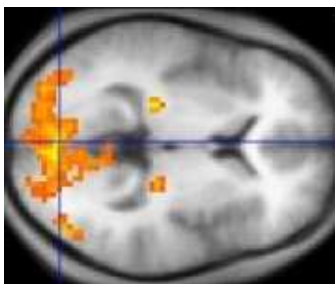


Figure 11: An fMRI image, showing active brain areas. (*biologydaily.com*).

Functional magnetic resonance imaging (fMRI) has emerged over the past decade as an indispensable technique for understanding the brain. Hemoglobin molecules change from diamagnetic to paramagnetic when deoxygenated, allowing fMRI to track changes in blood oxygenation. The fMRI blood oxygenation level-dependent (BOLD) signal shows which areas of the brain are most active at any given time, as illustrated in Fig. 11. (This is a very complex procedure: the signal is small and must be ‘deconvolved’ to produce ‘activity maps’ like that of Fig. 11.) Two and three-dimensional images can be reconstructed with good spatial resolution (on the order of 1 mm³), but data collection of each image takes 1-2 seconds. This temporal resolution is insufficient to track neural activity during rapid behaviors, but it can provide useful data for cognitive tasks that extend over several seconds (e.g. working memory). (Magnetoencephalography (MEG) offers much better temporal resolution, and spatial resolution close to that of fMRI.) Subjects can be given tasks while in the scanner, and brain areas showing increased activity during the task identified. These presumably play some part in the ‘circuit’ (assembly of areas) that executes the task. This technique has led to an explosion in the understanding of different brain regions, allowing substantial

refinement of Brodmann's divisions.

Electroencephalograms (EEGs) are collected from multiple scalp electrodes, much like electrocardiograms from the chest, and they offer excellent temporal resolution, but it is difficult to localize which brain areas are responsible for the signals detected at the scalp surface. This requires the solution of a hard electro-magnetic *inverse problems* like those used to interpret acoustic signals during prospecting for oil and minerals. (More beautiful mathematics, but not for this course.)

Behavioral experiments are also important, especially in cognitive psychology, even though they do not directly probe the anatomy and physiology of brain tissue. First of all, without observation and quantification of behavior during voltage recordings and fMRI scanning, the recorded data cannot be interpreted. In addition, behavioral data can be fit to 'high-level' abstracted models of neural systems, and important descriptors of the neural system thereby inferred. Behavioral data includes distributions of reaction times to stimuli, and decision times and error rates on cognitive choice tasks. In §6 we will discuss behavioral measures of this type, and describe some simple models that may link them with underlying neural processes.

2 Mathematical Tools: ODEs, Numerical Methods and Dynamical Systems (≈ 2.5 weeks)

In this section we provide a survival course in basic methods and ideas from the qualitative theory of ordinary differential equations (ODEs) and dynamical systems theory. Background material on linear ODEs can be found in textbooks such as [20], and on nonlinear systems in [100], and relevant ideas and results are also reviewed in texts on mathematical neuroscience, including [231] and [61, Chap. 3]. More advanced treatments are provided by [8] and [89]. Here we will introduce basic ideas that build on knowledge of multivariable calculus and linear algebra, and we will illustrate the theory with neurobiological examples.

The mathematical objects of interest are systems of first order ODEs, which may be written in compact vector notation as:

$$\frac{d}{dt}\mathbf{x} \stackrel{\text{def}}{=} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}); \quad \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad (2.0.1)$$

where \mathbf{x}_0 is an initial condition, and

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}. \quad (2.0.2)$$

This is sometimes called an ODEIVP (initial value problem) to distinguish it from boundary value problems (BVPs). The time-dependent quantities $x_j = x_j(t)$ are called *state variables* and they evolve in an n -dimensional *state space*, which we can usually take to be Euclidean space \mathbb{R}^n . Often the functions $f_j(\mathbf{x})$ depend upon *parameters* $\mu = (\mu_1, \dots, \mu_k)$, in which case we may write $\mathbf{f}(\mathbf{x}; \mu)$ or $\mathbf{f}_\mu(\mathbf{x})$, to indicate the different status of \mathbf{x} and μ . Indeed, while the parameters may themselves change, they usually do so on a slower scale than \mathbf{x} , and so may be taken as constant for the

purposes of solving (2.0.1). The function $\mathbf{f}(\mathbf{x})$ (or $\mathbf{f}(\mathbf{x}; \boldsymbol{\mu})$ or $\mathbf{f}_{\boldsymbol{\mu}}(\mathbf{x})$) defines a *vectorfield* on \mathbb{R}^n , which may be thought of as a forest of arrows telling which way solutions of (2.0.1) must go at each point. One can think of it as the velocity field of an (n -dimensional) “fluid” that permeates state space, and we say that it *generates a flow* on the state space. See Fig. 12.

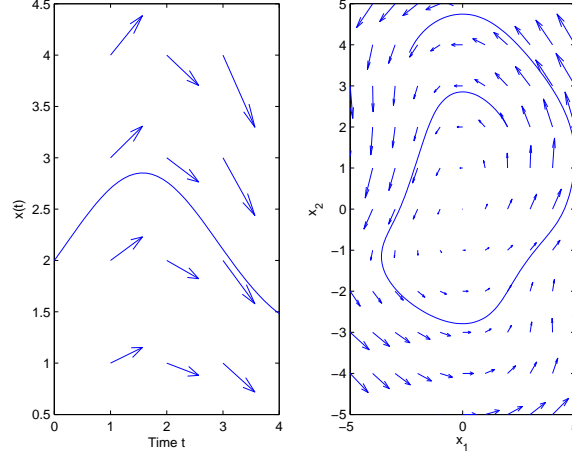


Figure 12: Vectorfields and some flowlines (solutions). (a) A non-autonomous scalar system $\dot{x} = f(x, t)$; (b) an autonomous two-dimensional system.

In general, flow maps are nonlinear, and while we cannot give explicit formulae like (2.1.5-2.1.6) below, the *local existence-uniqueness theorem* [100] guarantees that the nonlinear ODE (2.0.1) generates a *flow map*:

$$\mathbf{x}(t, \mathbf{x}_0) = \boldsymbol{\phi}_t(\mathbf{x}_0) \quad \text{or} \quad \boldsymbol{\phi}_t : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad (2.0.3)$$

at least for short times t , provided that the functions $f_j(\mathbf{x})$ (or, collectively, $\mathbf{f}(\mathbf{x})$) are smooth. This follows from assembling the unique solutions of (2.0.1) for each initial condition $\mathbf{x}(0) = \mathbf{x}_0$, which exist for finite time intervals around $t = 0$. Ideally, we would like to find sets of solutions to appropriate ranges of initial conditions, and to study how they depend upon those initial states, and on the parameters. Alas, it is a sad fact that explicit solutions expressible in terms of known functions can be found for very few functions \mathbf{f} . Nonetheless, it is often relatively easy to find special solutions such as *fixed points*, also called *equilibria*: points at which $\mathbf{f}(\mathbf{x}^e) = \mathbf{0}$ and so \mathbf{x}^e remains unchanged.

2.1 Linearization and solving linear systems of ODEs

Suppose that $\mathbf{x} = \mathbf{x}^e$ is an *equilibrium* or *fixed point*, i.e. $\mathbf{f}(\mathbf{x}^e) = \mathbf{0}$ ($\Leftrightarrow f_j(x_1^e, \dots, x_n^e) = 0$, $1 \leq j \leq n$). So, if $\mathbf{x}_0 = \mathbf{x}^e$, we conclude that $\mathbf{x}(t) \equiv \mathbf{x}^e$. Let $\mathbf{x}(t) = \mathbf{x}^e + \boldsymbol{\xi}(t)$, where $\boldsymbol{\xi}(t) = (\xi_1, \dots, \xi_n)$ is a small perturbation: $|\boldsymbol{\xi}(t)| \ll 1$. Substitute into (2.0.1) and expand \mathbf{f} in a multivariable, vector-valued Taylor series⁴ to obtain

$$\dot{\mathbf{x}}^e + \dot{\boldsymbol{\xi}} = \mathbf{f}(\mathbf{x}^e + \boldsymbol{\xi}) = \mathbf{f}(\mathbf{x}^e) + D\mathbf{f}(\mathbf{x}^e)\boldsymbol{\xi} + \mathcal{O}(|\boldsymbol{\xi}|^2). \quad (2.1.1)$$

⁴Here and in general we assume that \mathbf{f} is sufficiently differentiable so that Taylor’s Theorem with remainder applies to each component.

In (2.1.1) $D\mathbf{f}(\mathbf{x}^e)$ denotes the $n \times n$ Jacobian matrix of partial derivatives $\left[\frac{\partial f_i}{\partial x_j}\right]$, evaluated at the fixed point \mathbf{x}^e , and the “Big Oh” notation $\mathcal{O}(|\boldsymbol{\xi}|^2)$ characterizes the magnitude of quadratic and higher order terms in the components ξ_1, \dots, ξ_n . Specifically, if $\mathbf{g}(\boldsymbol{\xi}) = \mathcal{O}(|\boldsymbol{\xi}|^p)$ then

$$\lim_{|\boldsymbol{\xi}| \rightarrow 0} \frac{g(\boldsymbol{\xi})}{|\boldsymbol{\xi}|^p} \leq k < \infty. \quad (2.1.2)$$

More generally, $g(x) = \mathcal{O}(|h(x)|)$ as $x \rightarrow x_0$ means that the quotient $g(x)/h(x)$ is bounded as $x \rightarrow x_0$. (We will occasionally also use the less precise \sim notation, as in $\sin x \sim x$ and $\sin x \sim x - x^3/3$.)

Returning to Eqn. (2.1.1), for small enough $|\boldsymbol{\xi}|$, the first order term $D\mathbf{f}(\mathbf{x}^e)\boldsymbol{\xi}$ dominates. Taking into account that $\dot{\mathbf{x}}^e$ and $\mathbf{f}(\mathbf{x}^e)$ vanish and ignoring the small term $\mathcal{O}(|\boldsymbol{\xi}|^2)$, we get the linear system

$$\dot{\boldsymbol{\xi}} = D\mathbf{f}(\mathbf{x}^e)\boldsymbol{\xi} \quad (2.1.3)$$

The constant-coefficient linear ODE system (2.1.3) is called the *linearization* of (2.0.1) at \mathbf{x}^e . It can be solved by standard methods, as outlined below.

The general solution of (2.1.3) or of any linear system of the form

$$\dot{\mathbf{y}} = \mathbf{B}\mathbf{y}, \quad (2.1.4)$$

where \mathbf{B} is an $n \times n$ matrix and $\mathbf{y} = (y_1, \dots, y_n)$ is an n -vector, is determined by the eigenvalues and eigenvectors of \mathbf{B} (or $D\mathbf{f}(\mathbf{x}^e)$) [20]. Specifically, if the matrix \mathbf{B} has n linearly-independent eigenvectors $\mathbf{v}_i, i = 1, \dots, n$ with eigenvalues λ_i , we can form an $n \times n$, real-valued, *fundamental solution matrix* $\mathbf{X}^e(t)$ with columns $\mathbf{v}_i e^{\lambda_i t}$. As in (2.1.6), if complex eigenvalues $\lambda_{\pm} = \alpha \pm i\beta$ occur, \mathbf{X}^e can still be chosen real by taking adjacent columns as $e^{\alpha t}[\mathbf{u} \cos(\beta t) - \mathbf{w} \sin(\beta t)]$ and $e^{\alpha t}[\mathbf{w} \cos(\beta t) + \mathbf{u} \sin(\beta t)]$, where $\mathbf{v} = \mathbf{u} \pm i\mathbf{w}$ is the complex conjugate pair of eigenvectors belonging to λ_{\pm} . This follows from the facts that, if λ is a real eigenvalue with eigenvector \mathbf{v} , then there is a solution to (2.1.4) of the form:

$$\mathbf{y}(t) = c\mathbf{v}e^{\lambda t}, \quad (2.1.5)$$

and if $\lambda = \alpha + i\beta$ is a complex conjugate pair with eigenvectors $\mathbf{v} = \mathbf{u} + i\mathbf{w}$ (where \mathbf{u}, \mathbf{w} are real) then

$$\mathbf{y}_1(t) = e^{\alpha t}(\mathbf{u} \cos \beta t - \mathbf{w} \sin \beta t) \quad \text{and} \quad \mathbf{y}_2(t) = e^{\alpha t}(\mathbf{u} \sin \beta t + \mathbf{w} \cos \beta t) \quad (2.1.6)$$

are two linearly-independent solutions.

A short calculation (which we will do on demand) shows that, no matter how the columns of $\mathbf{X}^e(t)$ are ordered, the solution of (2.1.4) with initial condition $\mathbf{y}(0) = \mathbf{y}_0$ is given by:

$$\mathbf{y}(t) = \boldsymbol{\Phi}(t)\mathbf{y}_0, \quad \text{where} \quad \boldsymbol{\Phi}(t) = \mathbf{X}^e(t)[\mathbf{X}^e(0)]^{-1}. \quad (2.1.7)$$

This special fundamental solution matrix $\boldsymbol{\Phi}^e(t)$, which satisfies $\boldsymbol{\Phi}(0) = \mathbf{I}$, is the *flow map* for (2.1.4). It is an explicit rule (formula) that tells us how the vectorfield evolves solutions forward: it defines a time-dependent flow map that advances solutions in the state space \mathbb{R}^n . The flow map (2.1.7) is linear, like the ODE that defines it.

Exercise 1. (*Linear algebra review*) Verify that the flow map (2.1.7) does solve (2.1.4). What hypotheses must the eigenvectors \mathbf{v}_i satisfy for the matrix $\mathbf{X}^e(t)$ to be invertible? For what sorts of matrices B do these hypothesis always hold?

Here we are concerned with *qualitative* properties rather than exact or “complete” solutions. In particular, in studying stability we want to know whether the size of solutions grows, stays constant, or shrinks as $t \rightarrow \infty$. This can usually be answered just by looking at the eigenvalues⁵. Indeed, *the real part* of λ (almost) determines stability. Since *any* solution of (2.1.4) can be written as a linear superposition of terms of the forms (2.1.5)-(2.1.6) (except in the case of multiple eigenvalues), we can deduce that

- If all eigenvalues of \mathbf{B} have strictly negative real parts, then $|\mathbf{y}(t)| \rightarrow 0$ as $t \rightarrow \infty$ for all solutions.
- If at least one eigenvalue of \mathbf{B} has a positive real part, then there is a solution $\mathbf{y}(t)$ with $|\mathbf{y}(t)| \rightarrow +\infty$ as $t \rightarrow \infty$.
- If some eigenvalues have zero real part with *distinct* imaginary parts, then the corresponding solutions oscillate and neither decay nor grow as $t \rightarrow \infty$.

Definition 1. The fixed point \mathbf{x}^e of (2.0.1) is **hyperbolic** or *non-degenerate*, if all the eigenvalues of $D\mathbf{f}(\mathbf{x}^e)$ have non-zero real parts.

To make use of all this we need a notion of stability; we will do this in a geometrical context, suitable for our picture of phase space. See Figs. 13-14 for illustrations.

Definition 2 (Liapunov stability). \mathbf{x}^e is a **stable** fixed point of (2.0.1) if for every neighborhood $U \ni \mathbf{x}^e$ there is a neighborhood $V \subseteq U$ such that every solution $\mathbf{x}(t)$ of (2.0.1) starting in V ($\mathbf{x}(0) \in V$) remains in U for all $t \geq 0$. If \mathbf{x}^e is not stable, it is **unstable**.

Definition 3 (Asymptotic stability). \mathbf{x}^e is **asymptotically stable** if it is stable and additionally V can be chosen such that $|\mathbf{x}(t) - \mathbf{x}^e| \rightarrow 0$ as $t \rightarrow \infty$ for all $\mathbf{x}(0) \in V$.

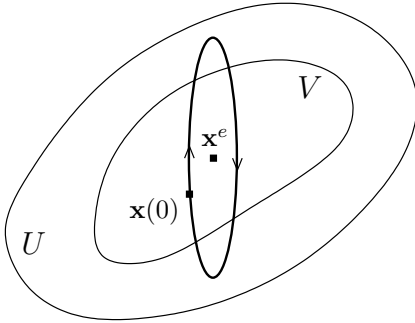


Figure 13: Stability.

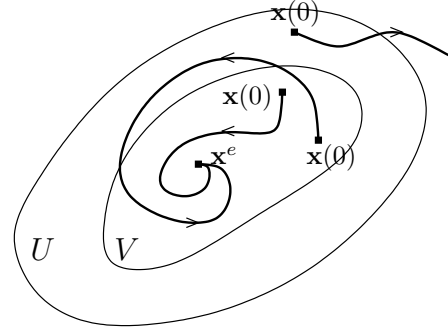


Figure 14: Asymptotic stability: a sink.

A fixed point that is Liapunov stable but not asymptotically stable is sometimes called *neutrally stable*. Equipped with these definitions and the linear analysis sketched above, and recognizing that the remainder terms ignored in passing from (2.1.1) to (2.1.4) can be made as small as we wish by selecting small neighborhood of \mathbf{x}^e , we can conclude that the stability type of the nonlinear system (2.0.1) under the flow ϕ_t is locally determined by the eigenvalues of $D\mathbf{f}(\mathbf{x}^e)$:

⁵Unless there are repeated eigenvalues (multiplicity ≥ 2). This is trickier – see [20, Ch. 7.7] or ask me.

Proposition 1. *If \mathbf{x}^e is a fixed point of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ and all the eigenvalues of $D\mathbf{f}(\mathbf{x}^e)$ have strictly negative real parts, then \mathbf{x}^e is asymptotically stable. If at least one eigenvalue has strictly positive real part, then \mathbf{x}^e is unstable.*

Further geometric information on the structure of solutions is given in §2.4.1.

Borrowing from fluid mechanics, we say that if all nearby solutions approach a fixed point (e.g. all eigenvalues have negative real parts), it is a *sink*; if all nearby solutions recede from it, it is a *source*, and if some approach and some recede, it is a *saddle point*. When the fixed point is surrounded by nested closed orbits, we call it a *center*.

We might hope to claim that stability (per Definition 2) holds even if (some) eigenvalues have zero real part, but the following counterexample crushes our hopes:

$$\dot{x} = \alpha x^3, \quad \alpha \neq 0. \quad (2.1.8)$$

Here $x = 0$ is the unique fixed point and the linearization at 0 is

$$\dot{\xi} = 3\alpha x^2|_{x=0} \xi = 0, \quad (2.1.9)$$

with solution $\xi(t) = \xi(0) = \text{const.}$, so certainly $x = 0$ is stable for (2.1.9). But the exact solution of the nonlinear ODE (2.1.8) may be found by separating variables:

$$\begin{aligned} \int_{x(0)}^{x(t)} \frac{dx}{x^3} &= \int \alpha dt \quad \Rightarrow \quad -\left(\frac{1}{2x(t)^2} - \frac{1}{2x(0)^2}\right) = \alpha t \\ \Rightarrow \quad x(t) &= \frac{x(0)}{\sqrt{1 - 2\alpha x(0)^2 t}}. \end{aligned}$$

We therefore deduce that

$$\left. \begin{array}{ll} |x(t)| \rightarrow \infty & \text{as } t \rightarrow \frac{1}{2\alpha x(0)^2} \text{ if } \alpha > 0 \quad (\text{blow up! instability}) \\ \text{but } |x(t)| \rightarrow 0 & \text{as } t \rightarrow \infty \text{ if } \alpha < 0 \quad (\text{asymptotic stability}) \end{array} \right\}.$$

The linearized system (2.1.9) is *degenerate* and the nonlinear “remainder terms,” ignored in our linearized analysis, determine the outcome in this case. Here it is obvious, at least in retrospect, that ignoring these terms is perilous, since while they are indeed $\mathcal{O}(\xi^2)$ (in fact, $\mathcal{O}(\xi^3)$), the linear $\mathcal{O}(\xi)$ term is identically zero! Moreover, global existence $\forall t \in \mathbb{R}$ fails for (2.1.8).

Example 1. *Consider the two-dimensional system*

$$\begin{aligned} \dot{x} &= y + \alpha(x^2 + y^2)x, \\ \dot{y} &= -x + \alpha(x^2 + y^2)y. \end{aligned}$$

Note that the linearization is simply a harmonic oscillator with eigenvalues $\pm i$. Is the equilibrium $(x, y) = (0, 0)$ of this system stable or unstable? To answer this, it is convenient to transform to polar coordinates $x = r \cos \theta$, $y = r \sin \theta$, which gives the uncoupled system:

$$\dot{r} = \alpha r^3, \quad \dot{\theta} = -1.$$

The first equation is as in the counterexample of (2.1.8) above, so we conclude: $\alpha > 0 \Rightarrow$ unstable; $\alpha = 0 \Rightarrow$ stable; $\alpha < 0 \Rightarrow$ asymptotically stable. The linearization gives no information if $\alpha \neq 0$.

How can we prove stability in such degenerate cases, in which one or more eigenvalues has zero real part? One method requires construction of a function which remains constant, or decreases, along solutions. For mechanical systems the total (kinetic plus potential) energy is often a good candidate.

2.2 Liapunov functions

Liapunov functions allow one to prove stability and even asymptotic stability in certain cases. We describe Liapunov's 'second method' or 'direct method' for ODEs of the form (2.0.1). Also see [231, Chap. 14].

Theorem 1. *Suppose that (2.0.1) has an isolated fixed point at $\mathbf{x} = \mathbf{0}$ (w.l.o.g. one can move a fixed point \mathbf{x}^e to $\mathbf{0}$ by letting $\mathbf{y} = \mathbf{x} - \mathbf{x}^e$). If there exists a differentiable function $V(\mathbf{x})$, which is positive definite and for which $\frac{dV}{dt} = \nabla V \cdot \mathbf{f}$ is negative definite on some domain $D \ni \mathbf{0}$, then $\mathbf{0}$ is asymptotically stable. If $\frac{dV}{dt}$ is negative semidefinite (i.e., $\frac{dV}{dt} = 0$ is allowed), then $\mathbf{0}$ is Liapunov-stable.*

Proof. (sketch) V is positive definite: $\left\{ \begin{array}{l} V(\mathbf{x}) > 0, \mathbf{x} \neq \mathbf{0} \\ V(\mathbf{x}) = 0, \mathbf{x} = \mathbf{0} \end{array} \right\} \Rightarrow$ the level sets of V are hyperspheres surrounding $\mathbf{x} = \mathbf{0}$.

\dot{V} is negative definite: $\left\{ \begin{array}{l} \dot{V}(\mathbf{x}) < 0, \mathbf{x} \neq \mathbf{0} \\ \dot{V}(\mathbf{x}) = 0, \mathbf{x} = \mathbf{0} \end{array} \right\} \Rightarrow$ solutions of (2.0.1) cross the level sets inwards converging on $\mathbf{x} = \mathbf{0}$. The level sets of V provide our neighborhoods U, V , each one crossed inwards by all solutions. Thus, as $\mathbf{x} \rightarrow \mathbf{0}$ we get an asymptotic stability.

\dot{V} is negative semidefinite: $\left\{ \begin{array}{l} \dot{V}(\mathbf{x}) \leq 0, \mathbf{x} \neq \mathbf{0} \\ \dot{V}(\mathbf{x}) = 0, \mathbf{x} = \mathbf{0} \end{array} \right\} \Rightarrow$ solutions of (2.0.1) either cross level sets inwards, or are confined to level sets. Again, we use the level sets of V for obtaining neighborhoods U, V of \mathbf{x} ; but cannot conclude that $\mathbf{0}$ is asymptotically stable. \square

Example 2. *Example 1 revisited. Let $V = (x^2 + y^2)/2$ and compute*

$$\dot{V} = x\dot{x} + y\dot{y} = x(y + \alpha(x^2 + y^2)x) + y(-x + \alpha(x^2 + y^2)y) = \alpha(x^2 + y^2)^2.$$

We see that, if $\alpha < 0$ $\dot{V} < 0$ for all $(x, y) \neq (0, 0)$, implying asymptotic stability, while for $\alpha > 0$ $\dot{V} > 0$ for all $(x, y) \neq (0, 0)$, implying instability.

The notions of stability in Definitions 2-3 may be generalized to non-constant orbits of ODEs (periodic, quasiperiodic, or non-periodic). First, some notation and preparatory definitions are needed. Let ϕ_t be the flow map of Eqn. (2.0.3) and let $\phi(\mathbf{x}) = \{\phi_t(\mathbf{x}) | t \geq 0\}$ denote the set of all points in the solution or orbit $\phi_t(\mathbf{x})$ based at \mathbf{x} .

Definition 4. *Two orbits $\phi(\mathbf{x})$ and $\phi(\hat{\mathbf{x}})$ are ϵ -close if there is a reparameterization of time $\hat{t}(t)$ (a smooth, monotonic function) such that $|\phi_t(\mathbf{x}) - \phi_{\hat{t}(t)}(\hat{\mathbf{x}})| < \epsilon$ for all $t \geq 0$.*

Neighboring periodic orbits in a conservative system may circulate at different speeds, as in the following example (in polar coordinates, cf. Example 1):

$$\dot{r} = 0, \quad \dot{\theta} = 1 + r^2,$$

where the period depends on amplitude: $T(r) = 2\pi/(1+r^2)$. Reparameterization of time allows for this by bringing neighboring solutions back into step. Monotonicity of $\hat{t}(t)$ implies that the orbits are not close merely as sets, but that after the change of timescale every pair of their values $\phi_t(\mathbf{x})$ and $\phi_{\hat{t}}(\hat{\mathbf{x}})$ at succeeding times are close.

Definition 5. A solution $\phi_t(\mathbf{x})$ is **orbitally stable** if for every $\epsilon > 0$ there is a neighborhood $V \ni \mathbf{x}$ such that, for all $\hat{\mathbf{x}} \in V$, the sets $\phi(\mathbf{x})$ and $\phi(\hat{\mathbf{x}})$ are ϵ -close. If additionally V may be chosen such that for all $\hat{\mathbf{x}} \in V$, there exists a time shift $\tau(\hat{\mathbf{x}})$ so that $|\phi_t(\mathbf{x}) - \phi_{t-\tau(\hat{\mathbf{x}})}(\hat{\mathbf{x}})| \rightarrow 0$ as $t \rightarrow \infty$, then $\phi_t(\mathbf{x})$ is **orbitally asymptotically stable**.

See Figs. 15-16, which show (segments of) the orbits $\phi_t(\mathbf{x})$ and their neighborhoods V , each of which contains a second orbit based at $\hat{\mathbf{x}}$.

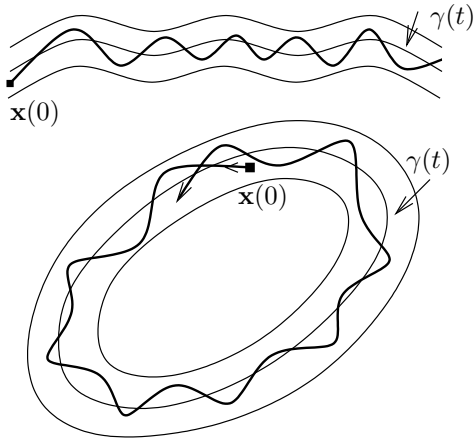


Figure 15: Orbital stability

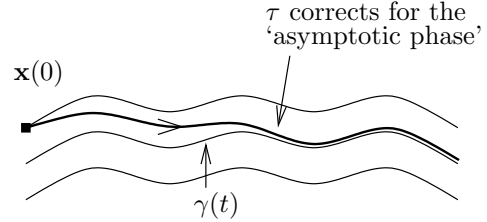


Figure 16: Orbital asymptotic stability

2.3 Numerical methods

As we have remarked, very few ODEs admit explicit solutions in terms of “nice” known functions (trigonometric, exponential, polynomial, etc.) We therefore resort to numerical solution, happily calling up Matlab or similar software. But can we believe what it tells us? This section outlines the main *mathematical* ideas behind numerical integration of ODEs, starting with the simplest such method, which we treat in some detail. For (much) more information and background, go to a text on numerical methods, such as [206]. **CITE MORE BASIC ODE NA REF??**

2.3.1 Euler’s method

Forward Euler method: Consider the general scalar (one-phase-space dimension) ODE

$$\dot{x} = f(x, t), \quad x(t_0) = x_0, \quad x \in \mathbb{R}. \quad (2.3.1)$$

Euler's method is the simplest, but least efficient numerical integrator in terms of the number of steps needed to propagate solutions for times $t \in [t_0, t_0 + T]$ with given accuracy. It discretizes the derivative $\dot{x} \approx (x_{n+1} - x_n)/\Delta t$ to yield the difference equation:

$$x_{n+1} = x_n + \Delta t f(x_n, t_n), \quad (2.3.2)$$

with $x_0 = x(t_0)$ (naturally), and $t_n = t_0 + n\Delta t$. This is the *forward* Euler method (see below for backward Euler). For $x \in \mathbb{R}^n$, we simply do the above for each component:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \Delta t \mathbf{f}(\mathbf{x}_n, t_n), \quad (2.3.3a)$$

$$x_{n+1}^1 = x_n^1 + \Delta t f_1(x_n^1, x_n^2, \dots, x_n^N, t_n), \quad (2.3.3b)$$

$$x_{n+1}^2 = x_n^2 + \Delta t f_2(x_n^1, x_n^2, \dots, x_n^N, t_n), \quad (2.3.3c)$$

$$\vdots$$

$$x_{n+1}^N = x_n^N + \Delta t f_N(x_n^1, x_n^2, \dots, x_n^N, t_n), \quad (2.3.3d)$$

Example 3. Consider the harmonic oscillator

$$\ddot{x} + x = 0 \quad \text{or} \quad \dot{x} = y, \quad \dot{y} = -x, \quad \text{with } x(0) = x_0, \quad y(0) = y_0,$$

the exact solution of which can be expressed via a linear map that is a rotation matrix:

$$\begin{aligned} x(t) &= x_0 \cos(t) + y_0 \sin(t), \\ y(t) &= -x_0 \sin(t) + y_0 \cos(t), \\ \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= \begin{bmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{bmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \\ \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} &= X(t) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}. \end{aligned}$$

This is an example of the general fundamental solution matrix procedure sketched in §2.1. Euler's method approximates the harmonic oscillator as

$$\begin{aligned} x_{n+1} &= x_n + \Delta t y_n, \\ y_{n+1} &= y_n - \Delta t x_n, \end{aligned}$$

which is a linear transformation:

$$\begin{pmatrix} x \\ y \end{pmatrix}_{n+1} = [A] \begin{pmatrix} x \\ y \end{pmatrix}_n, \quad A = \begin{bmatrix} 1 & \Delta t \\ -\Delta t & 1 \end{bmatrix}$$

and $\det[A] = 1 + (\Delta t)^2 > 1$, so **areas are expanded**. This approximation to the solution matrix is **qualitatively incorrect**, since the exact rotation preserves areas, because $\det[X(t)] = \cos^2 t + \sin^2 t \equiv 1$. In the $(x, y = \dot{x})$ phase plane, the exact solutions form circles, but the Euler method steps continually move out across the circles, each finite step being a segment of a tangent vector to one of the circles, as seen in Fig. 17. The numerical scheme has destroyed the conservative (Hamiltonian) structure of the original problem.

Moral: we can't expect numerical schemes to preserve constants of motion, constraints, or the overall structure of equations (but **symplectic integrators** do preserve Hamiltonian structures).

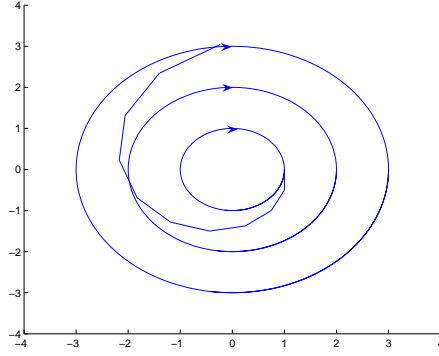


Figure 17: Exact and Euler-approximated solutions to the harmonic oscillator $\ddot{x} + x = 0$. On the circles, the total energy $\frac{x^2}{2} + \frac{y^2}{2} = \frac{x^2 + \dot{x}^2}{2} = \text{const.}$ A large step size has been taken to emphasize errors.

Back in one dimension, the Euler method is also called the tangent line method, since at each iteration we take the step

$$x_{n+1} - x_n = \Delta t f(x_n, t_n) \quad (2.3.4)$$

in the direction of the tangent to a “true” solution, as in Fig. 18.

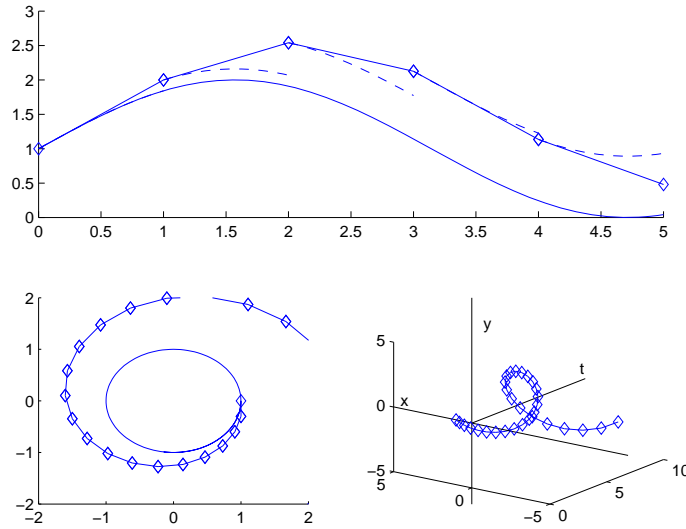


Figure 18: Top: the forward Euler or tangent-line method; solid and dashed curves are exact solutions, segments with diamonds are Euler steps. Bottom: solutions of Example 3 projected onto the (x, y) plane and plotted in three dimensional (x, y, t) -space.

Note that Euler’s method essentially computes a truncated Taylor series of the solution. Letting $\phi(t)$ denote the true (exact) solution based at $\phi(t_0) = x_0$, we have

$$\phi(t) = \phi(t_0) + \phi'(t_0)(t - t_0) + \frac{\phi''(t_0)}{2!}(t - t_0)^2 + \mathcal{O}((t - t_0)^3), \quad (2.3.5)$$

with $\phi(t_0) = x_0$ and $\phi'(t_0) = f(x_0, t_0)$. Setting $t = t_1$ and $t - t_0 = t_1 - t_0 = \Delta t$ (uniform step sizes), (2.3.5) gives

$$\phi(t_1) = x_0 + f(x_0, t_0)\Delta t + \frac{\phi''(t_0)}{2!}\Delta t^2 + \mathcal{O}(\Delta t^3). \quad (2.3.6)$$

So $x_1 = x_0 + f(x_0, t_0)\Delta t$ approximates the true solution $\phi(t_1) = \phi(t_0 + \Delta t)$ with an error of size $|x_1 - \phi(t_1)| = \mathcal{O}(\Delta t^2)$. Similarly, for any successive step,

$$\phi(t_{n+1}) = \phi(t_n) + \phi'(t_n)\Delta t + \mathcal{O}(\Delta t^2),$$

so that, provided $|x_n - \phi(t_n)| = \mathcal{O}(\Delta t^2)$, we also have

$$x_{n+1} = x_n + f(x_n, t_n)\Delta t = \phi(t_{n+1}) + \mathcal{O}(\Delta t^2) : \quad (2.3.7)$$

at each step we get an error of $\mathcal{O}(\Delta t^2)$. We shall shortly see how these errors accumulate.

Backward Euler method: Now for the *backward* Euler method. For the forward Euler method, as (2.3.4) shows, we're approximating the time derivative $\dot{x} = f(x, t)$ by a forward difference quotient: $\frac{x_{n+1} - x_n}{\Delta t} = f(x_n, t_n)$. We can also approximate it by a backward difference quotient

$$\frac{x_{n+1} - x_n}{\Delta t} = f(x_{n+1}, t_{n+1}) \quad \text{or} \quad x_{n+1} = x_n + \Delta t f(x_{n+1}, t_{n+1}). \quad (2.3.8)$$

This formula defines the backward Euler method. Note that *in general*, (2.3.8) must be solved iteratively: we have to find the point x_{n+1} that we would have arrived at, leaving *from* x_n in the direction tangent to the vector field *at* x_{n+1} . The forward Euler method is *explicit* – we just plug in x_n and compute – but backward Euler is *implicit*: x_{n+1} must be solved for.

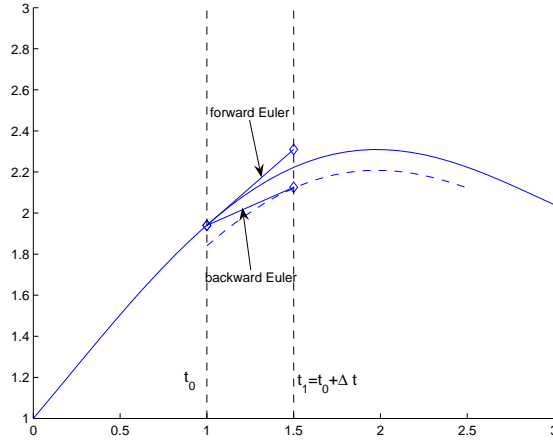


Figure 19: The forward and backward Euler methods compared. The solid and dashed curves denote exact solutions.

Example 4. The linear ODE $\dot{x} = ax$, $x(0) = x_0$ is a case that we can solve explicitly for both methods. Forward Euler leads to

$$x_{n+1} = x_n + \Delta t(ax_n), \quad \text{with initial condition } x_0 \quad (2.3.9a)$$

$$\Rightarrow x_n = x_0(1 + a\Delta t)^n. \quad (2.3.9b)$$

For backward Euler we have

$$x_1 = x_0 + \Delta t(ax_1) \Rightarrow (1 - a\Delta t)x_1 = x_0 \Rightarrow x_1 = x_0(1 - a\Delta t)^{-1}, \quad (2.3.10a)$$

$$x_2 = x_1 + \Delta t(ax_2) \Rightarrow (1 - a\Delta t)x_2 = x_1 = x_0(1 - a\Delta t)^{-1} \Rightarrow x_2 = x_0(1 - a\Delta t)^{-2} \dots \quad (2.3.10b)$$

$$x_n = x_0(1 - a\Delta t)^{-n}. \quad (2.3.10c)$$

Both methods agree up to $\mathcal{O}(\Delta t)$ [$x_n = x_0(1 + na\Delta t + \dots)$], but thereafter they disagree. Both are accurate up to $\mathcal{O}(\Delta t)$, with $\mathcal{O}(\Delta t^2)$ error per step. This is true in general, and not just for the exponential solution of $\dot{x} = ax$.

As we've pointed out, the backward Euler formula (2.3.8) may not be explicitly soluble, so we may have to use, e.g., Newton's method to find x_{n+1} at each step: an iterative loop within an iterative loop. (See §2.3.4 for Newton's method). So why be backward, if forward Euler is much simpler? The answer has to do with *stability*. Example 3 already shows that forward Euler can turn a (Liapunov) stable system, the harmonic oscillator, into an unstable growing spiral. More generally, even for an asymptotically stable scalar equation whose solutions all decay to fixed points or approach particular solutions $\bar{x}(t)$ as $t \rightarrow \infty$, the forward Euler method can give *growing* solutions if the step size Δt is too large. In contrast, the backward Euler method is *unconditionally stable*.

Example 5. Consider $\dot{x} = -ax$, $x(0) = x_0$, $a > 0$, for which the exact solution $x(t) = x_0 e^{-at}$ is a decaying exponential. Forward Euler gives:

$$x_{n+1} = x_n - a\Delta t x_n = (1 - a\Delta t)x_n \quad (2.3.11)$$

$$\Rightarrow x_n = (1 - a\Delta t)^n x_0 \rightarrow 0 \text{ if and only if } |1 - a\Delta t| < 1.$$

If we pick $a\Delta t < 2$, then $|1 - a\Delta t| < 1$ and our numerical scheme reflects the stability and 'contraction' of solutions inherent in the equation. But if $a\Delta t > 2$ then $1 - a\Delta t < -1$ and $(1 - a\Delta t)^n \rightarrow \infty$ in a violent oscillatory manner. Hence $\Delta t < \frac{2}{a}$ is a necessary and sufficient condition for forward Euler to preserve the stability inherent in this example. We may appeal to linearization of $\dot{x} = f(x, t)$ near specific solutions $\bar{x}(t)$ to generalize this argument. Conclusion: the forward Euler method is only conditionally stable.

The backward Euler's behavior is different: it is **unconditionally stable**:

$$\begin{aligned} x_{n+1} &= x_n - a\Delta t x_{n+1} \Rightarrow x_{n+1} = \frac{x_n}{1 + a\Delta t} \\ \Rightarrow x_n &= \frac{x_0}{(1 + a\Delta t)^n} \rightarrow 0, \quad \forall \Delta t > 0 \text{ if } a > 0 !! \end{aligned} \quad (2.3.12)$$

In general, implicit or backward schemes have better stability properties: they look ahead, and so do better in reproducing forward asymptotic behavior.

Stability: We now take a *geometric* view of oscillatory instability of forward Euler and unconditional stability of backward Euler (again for the exponential decay equation). As we have seen, the forward Euler method applied to $\dot{x} = -ax$ gives the approximate discrete solution

$x_n = x_0(1 - a\Delta t)^n$, so that $a\Delta t > 2 \Rightarrow$ instability. More specifically, with $a = 1$, we observe the behaviors shown in Fig. 20:

$$\begin{aligned} \Delta t = \frac{1}{2} \quad x_n &= x_0 \left(\frac{1}{2}\right)^n, \quad \text{monotonic decay} \rightarrow 0; \\ \Delta t = \frac{3}{2} \quad x_n &= x_0 \left(-\frac{1}{2}\right)^n, \quad \text{oscillatory decay} \rightarrow 0 : \text{ correct as } t \rightarrow \infty, \text{ but a bad approach;} \\ \Delta t = 2 \quad x_n &= x_0(-1)^n, \quad \text{period 2 orbit } x_0 \rightarrow -x_0 \rightarrow x_0 \rightarrow \cdots : \text{ entirely incorrect;} \\ \Delta t = \frac{5}{2} \quad x_n &= x_0 \left(-\frac{3}{2}\right)^n, \quad \text{oscillatory growth, } x_0 \rightarrow -\frac{3x_0}{2} \rightarrow \frac{9x_0}{4} \rightarrow \cdots : \text{ a disaster!} \end{aligned}$$

Note that the stability condition $\Delta t < \frac{2}{a}$ becomes *more* stringent, requiring smaller Δt , for *larger* a . Paradoxically, very stable systems are harder to integrate! There are special methods for these *stiff* systems.

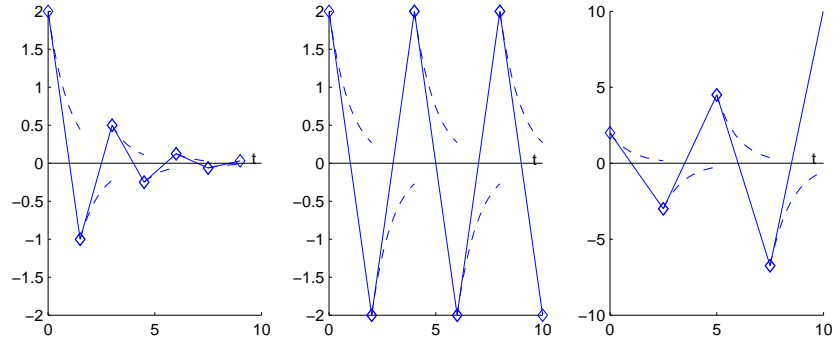


Figure 20: Oscillatory instability of forward Euler for $a = 1$, and $\Delta t = 1.5, 2$, and 2.5 .

In contrast, for the backward Euler algorithm $x_{n+1} = (1 + a\Delta t)^{-1}x_n$, monotonic decay persists as Δt increases, although the decay rate is increasingly underestimated: see Fig. 21.

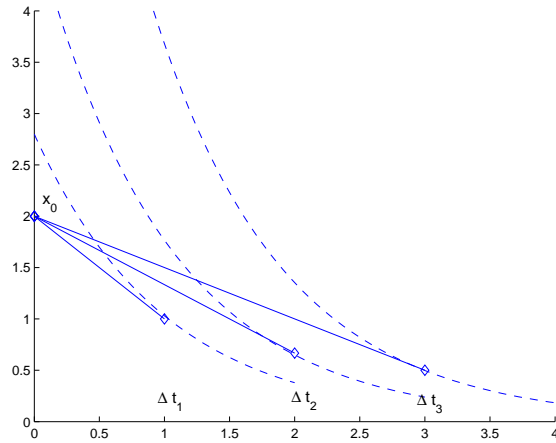


Figure 21: Unconditional stability of backward Euler for step sizes $\Delta t = 1, 2$, and 3 .

Error Analysis: We have already noted that Euler's method (forward or backward) produces an error $\mathcal{O}(\Delta t^2)$ at each step (Eq. (2.3.7)). As earlier, letting $\phi(t)$ denote the exact solution and x_n denote the (forward) Euler approximation, we compare the Taylor series expansion of ϕ with x_n . Recalling that $\phi'(t_n) = f(\phi(t_n), t_n)$ and using Taylor's theorem with remainder, we have the *exact* expressions

$$\phi(t_{n+1}) = \phi(t_n) + \phi'(t_n)\Delta t + \frac{\phi''(\tilde{t}_n)}{2!}\Delta t^2, \quad \tilde{t}_n \in [t_n, t_{n+1}], \quad (2.3.13a)$$

$$x_{n+1} = x_n + f(x_n, t_n)\Delta t. \quad (2.3.13b)$$

Denoting the *error* at step n as $e_n \stackrel{\text{def}}{=} |x_n - \phi(t_n)|$ and subtracting (2.3.13a) and (2.3.13b) leads to

$$\begin{aligned} e_{n+1} &= |x_{n+1} - \phi(t_{n+1})| = \left| x_n + f(x_n, t_n)\Delta t - \phi(t_n) - f(\phi(t_n), t_n)\Delta t - \phi''(\tilde{t}_n)\frac{\Delta t^2}{2} \right| \\ \Rightarrow e_{n+1} &\leq |x_n - \phi(t_n)| + |f(x_n, t_n) - f(\phi(t_n), t_n)|\Delta t + \frac{1}{2}|\phi''(\tilde{t}_n)|\Delta t^2. \end{aligned} \quad (2.3.14)$$

Stepwise error: Suppose $x_n = \phi(t_n)$ (we start at the right place), then (2.3.14) $\Rightarrow e_{n+1} \leq \frac{1}{2}|\phi''(\tilde{t}_n)|\Delta t^2$, so the error is $\mathcal{O}(\Delta t^2)$ per step, as we've already seen. In fact we can state this in terms of a uniform bound on the second derivative of solutions: $e_n \leq M\frac{\Delta t^2}{2}$, where $M = \max_{t \in [t_0, t_0+T]} |\phi''(t)|$.

Global accumulated error: Now pick a stepsize Δt and take N steps for a total elapsed time of $T = N\Delta t$. We suppose that T is fixed and ask: "What is the largest total accumulated error at time $t = T$?" To answer this we let K be the Lipschitz constant⁶ of $f(x, t)$ with respect to $x \in U$, and assume that it holds for all $t \in [0, T]$, where U is a bounded set in which the solution $\phi(t)$ remains for $t \in [t_0, t_0 + T]$. Then (2.3.14) implies that

$$\begin{aligned} e_{n+1} &\leq e_n + K e_n \Delta t + M \frac{\Delta t^2}{2}, \quad M = \max_{t \in [t_0, t_0+T]} |\phi''(t)|, \\ e_{n+1} &\leq (1 + K\Delta t)e_n + \frac{M\Delta t^2}{2}, \quad e_0 = 0 \quad (\text{start at the right place}). \end{aligned}$$

Setting $A = (1 + K\Delta t)$ and $B = \frac{M\Delta t^2}{2}$, we solve this difference inequality

$$e_{n+1} \leq A e_n + B, \quad \text{with } e_0 = 0 \quad (2.3.15)$$

as follows:

$$\begin{aligned} e_1 &\leq A e_0 + B = B, \\ e_2 &\leq A e_1 + B \leq AB + B = (A + 1)B, \\ e_3 &\leq A e_2 + B \leq A(A + 1)B + B = (A^2 + A + 1)B, \\ &\vdots \\ e_N &\leq (A^{N-1} + A^{N-2} + \cdots + A + 1)B = B \sum_{j=0}^{N-1} A^j = B \left(\frac{A^N - 1}{A - 1} \right). \end{aligned}$$

⁶If $g(x)$ has Lipschitz constant K then $|g(x) - g(y)| \leq K|x - y|$. If g is differentiable then K is the maximum value of g' , but Lipschitz constants can also be defined for continuous non-differentiable functions like $g(x) = |x|$.

$$\Rightarrow e_N \leq \frac{M\Delta t^2}{2} \left[\frac{(1 + K\Delta t)^N - 1}{K\Delta t} \right] \leq \frac{M\Delta t^2}{2} \left[\frac{e^{NK\Delta t} - 1}{K\Delta t} \right]; \quad (2.3.16)$$

the last inequality uses $(1 + \alpha)^N \leq 1 + \alpha N + \frac{\alpha^2 N^2}{2!} + \dots = e^{\alpha N}$. We conclude that

$$e_N \leq \frac{M\Delta t}{2K} [e^{KN\Delta t} - 1] = \frac{M\Delta t}{2K} [e^{KT} - 1] \stackrel{\text{def}}{=} \mathcal{K}\Delta t, \quad (2.3.17)$$

with a constant \mathcal{K} that depends on K , M , and T , but *not* on N . Hence the global accumulated error is $\mathcal{O}(\Delta t)$. Euler's scheme is therefore called a *first-order method*: to double the accuracy at a fixed time T , we must halve Δt and thus double the number of steps N .

2.3.2 The second-order Euler-Heun method

Let $\phi(t)$ be an exact solution of $\dot{x} = f(x, t)$ and consider the formula derived integrating the ODE:

$$\phi(t_{n+1}) = \phi(t_n) + \int_{t_n}^{t_{n+1}} f(\phi(s), s) ds. \quad (2.3.18)$$

(This is an integral equation for the solution we seek, but of course we can't solve it for general nonlinear functions f !) Euler's (forward or backward) method approximates the integral of (2.3.18) simply by the product of steplength $(t_{n+1} - t_n)$ and the value of f either at t_n or t_{n+1} . A better approximation to this integral should therefore improve the estimate. As Fig. 22 shows, Euler uses a rectangular (cf. Riemann sum) approximation. The second-order Euler-Heun method uses a trapezoidal approximation:

$$\int_{t_n}^{t_{n+1}} f(\phi(s), s) ds \approx \frac{t_{n+1} - t_n}{2} [f(\phi(t_n), t_n) + f(\phi(t_{n+1}), t_{n+1})]. \quad (2.3.19)$$

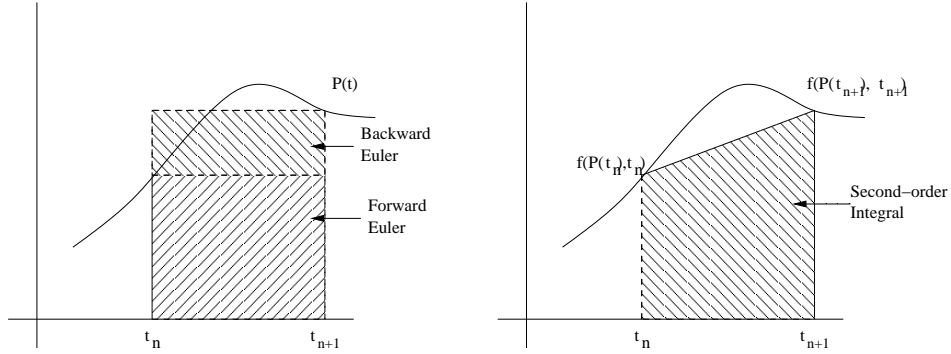


Figure 22: First-order and second-order approximations to the integral in (2.3.18).

Of course, we don't know $\phi(t)$ and so cannot calculate this expression explicitly, but we do our best by writing $x_n \approx \phi(t_n)$, $x_{n+1} \approx \phi(t_{n+1})$, which gives:

$$x_{n+1} = x_n + \frac{\Delta t}{2} [f(x_n, t_n) + f(x_{n+1}, t_{n+1})], \quad \Delta t = t_{n+1} - t_n. \quad (2.3.20)$$

This is still an implicit method (x_{n+1} appears on RHS), but we may make it explicit at the cost of another approximation, by estimating $x_{n+1} = x_n + \Delta t f(x_n, t_n)$ from first-order Euler:

$$x_{n+1} = x_n + \frac{\Delta t}{2} [f(x_n, t_n) + f(x_n + \Delta t f(x_n, t_n), t_{n+1})]. \quad (2.3.21)$$

This defines the improved Euler or Heun method. The stepwise error is $\mathcal{O}(\Delta t^3)$ and the global accumulated error is $\mathcal{O}(\Delta t^2)$ (not proved here). Halving the step size now improves accuracy by a factor of 4!

2.3.3 The Runge-Kutta method

The Runge-Kutta (RK) method is the basis of many ‘integration packages’, including ones called by Matlab. Here’s the recipe for a fixed step size RK method for the scalar ODE $\dot{x} = f(x, t)$:

$$x_{n+1} = x_n + \frac{\Delta t}{6} [k_n^1 + 2k_n^2 + 2k_n^3 + k_n^4], \text{ where} \quad (2.3.22a)$$

$$k_n^1 = f(x_n, t_n), \quad (2.3.22b)$$

$$k_n^2 = f(x_n + \frac{\Delta t k_n^1}{2}, t_n + \frac{\Delta t}{2}), \quad (2.3.22c)$$

$$k_n^3 = f(x_n + \frac{\Delta t k_n^2}{2}, t_n + \frac{\Delta t}{2}), \quad (2.3.22d)$$

$$k_n^4 = f(x_n + \Delta t k_n^3, t_n + \Delta t). \quad (2.3.22e)$$

This does a much better job of approximating the integral (2.3.18) by averaging the slope of the vector field over its values at *four* points in the step, as pictured in Fig. 23.

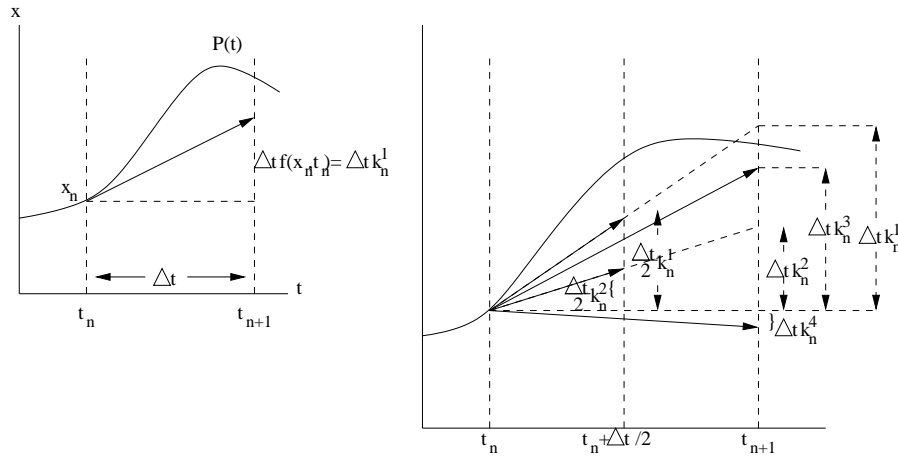


Figure 23: Comparison of 1 point average forward-Euler and 4 point average Runge-Kutta.

The resulting stepwise error is $\mathcal{O}(\Delta t^5)$ and the global accumulated error is $\mathcal{O}(\Delta t^4)$, so this is a fourth order method, abbreviated as RK4 (there is also a fifth order RK method). More calculations are required per step, but far fewer steps are needed to obtain the same accuracy at time T . For example, to obtain accuracy of $\mathcal{O}(10^{-6})$ at $T = 1$ for first-order Euler, we need $\Delta t \sim 10^{-6}$ and

thus $\mathcal{O}(10^6)$ steps. For second-order Euler/Heun: $\Delta t \sim 10^{-3}$ and $\mathcal{O}(10^3)$ steps, but for fourth-order Runge-Kutta, $\Delta t \approx \frac{1}{30}$ requires only $\mathcal{O}(10^{\frac{6}{4}}) = \mathcal{O}(\sqrt{1000}) = \mathcal{O}(30)$ steps. The extension of second-order Euler-Heun and RK4 to *systems* is done component-wise as in (2.3.3).

Here we have considered only fixed step methods, for which $\Delta t = t_{n+1} - t_n, \forall n$. Clearly, we could take larger steps in regions in which $f(x, t)$ varies little (or ‘slowly’), so that the stepwise error constant M is small, and, to maintain the same absolute accuracy, take shorter steps in regions in which f changes rapidly. Such variable stepsize methods can also reduce computational effort. At each step, the error is estimated and larger or shorter steps are used, sometimes in a predictor/corrector strategy. The Matlab routine `ode45()` uses such a strategy. However, see [202] for an interesting example of how it can fail.

2.3.4 Newton’s method for finding zeros

Finding the zeros of a function seems a simpler problem than solving an ODE: one must only approximate some numbers, not a whole function. However, it also requires an iterative approach. Suppose we want to solve $f(x) = 0$, with $f \in C^1$ (at least once-differentiable). This can be an embarrassing question for a mathematician. For example, there is no closed form formula for the roots of a polynomial of n th order for $n \geq 5$, and he or she must resort to a numerical method.

We define the iterative process:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \text{where } f'(x) = \frac{df}{dx}(x). \quad (2.3.23)$$

Proposition 2. *For almost all x_0 , the sequence generated by (2.3.23) $\{x_n\}_{n=0}^\infty \rightarrow \bar{x}$, where $f(\bar{x}) = 0$, i.e. the sequence converges on one of the (possibly many) solutions of $f(x) = 0$.*

Proof. (Sketch) Provided $\frac{f(\bar{x})}{f'(\bar{x})} = 0$ the limit equation $\bar{x} = \bar{x} - \frac{f(\bar{x})}{f'(\bar{x})}$ is clearly satisfied, and so we can expect a well-defined limit point provided that the zero is *simple*: $f'(\bar{x}) \neq 0$. This is true for almost all zeros, since otherwise **two** conditions $f(\bar{x}) = f'(\bar{x}) = 0$ must hold simultaneously. Moreover, the fixed point \bar{x} of the iterated map (2.3.23) is locally asymptotically stable, as we see by linearizing (2.3.23) at \bar{x} . Letting $x = \bar{x} + y$ and expanding in a Taylor series as we did when linearizing ODEs, (2.3.23) becomes

$$\begin{aligned} \bar{x} + y_{n+1} &= \bar{x} + y_n - \frac{f(\bar{x} + y_n)}{f'(\bar{x} + y_n)}, \\ \Rightarrow y_{n+1} &= y_n - \left(\frac{f}{f'} \right)'(\bar{x}) y_n + \mathcal{O}(y_n^2) = y_n - \left[\frac{f'f' - ff''}{(f')^2} \right] \Big|_{x=\bar{x}} y_n + \mathcal{O}(y_n^2). \end{aligned}$$

Truncating at the linear terms, we have

$$y_{n+1} = \left[1 - \frac{f'(\bar{x})^2 - f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} \right] y_n = 0, \quad (2.3.24)$$

since $f(\bar{x}) = 0$ and $\frac{f'(\bar{x})}{f'(\bar{x})} = 1$. So the eigenvalue $\lambda = 0$ for this one-dimensional map, and we have superlinear or quadratic convergence. \square

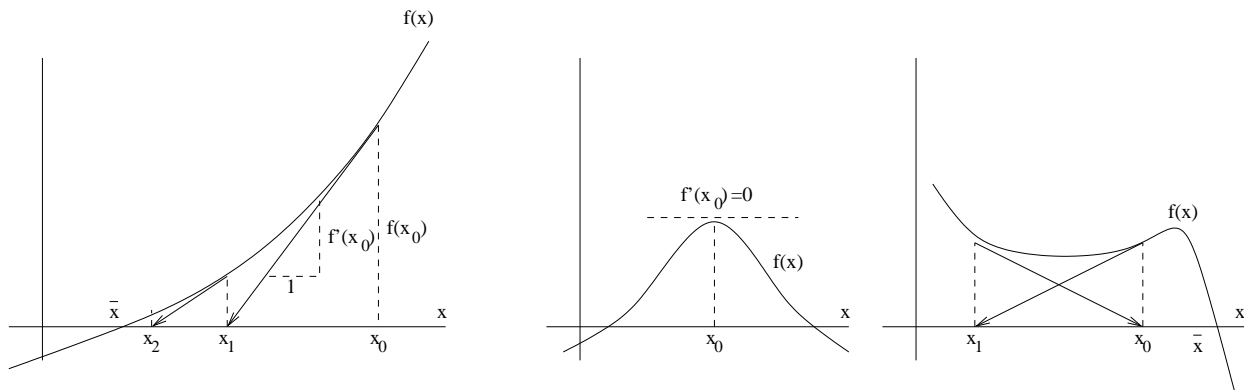


Figure 24: Geometric interpretation of Newton's method. a) Normal case, with basin of attraction and quadratic convergence rate. b) A 'bad' initial guess x_0 for which $f'(x_0) = 0$. c) Another bad case: x_0 lies in a periodic cycle.

There is a geometric interpretation to Newton's method as can be seen in Fig. 24. In a), all initial data is in the 'basin of attraction' of \bar{x} , and approaches \bar{x} under iteration (2.3.23). In b), there is an example of a 'bad' x_0 , at which $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = \pm\infty$!! Finally, in c) we see x_0 in an (unstable) cycle. Consider small perturbations of these bad choices, and think about what might happen if $f(x)$ has multiple roots.

Numerical methods is a **HUGE** field! MAT 342, APC523 and other courses do a proper job on it. This brief taste is only an appetizer!

2.4 Geometric theory of dynamical systems

In his attempts to solve problems of celestial mechanics in the late 19th century, the great French mathematician Henri Poincaré thought deeply about nonlinear differential equations. His ideas, research papers, and books (e.g. [166, 167]) gave rise to the qualitative or geometric theory of dynamical systems. The following, partially taken from [108], provides an informal introduction to some of the main ideas. For background and deeper coverage of the main ideas, see [100], and for a more advanced treatment, see [89] (and take MAE541/APC571!).

Dynamical systems are more general objects than differential equations: they consist of a *phase* or *state space* M and a rule that specifies how points $\mathbf{x} \in M$ (system states) evolve as time progresses. Evolution can be continuous, as in ODEs, or discrete, as in iterated maps of the form:

$$\mathbf{x}_{n+1} = \mathbf{F}(\mathbf{x}_n), \quad (2.4.1)$$

examples of which include the numerical algorithms of §2.3. (In fact dynamical systems ideas can in turn be used to analyse numerical algorithms such as Newton's method for finding zeros [207].) Solutions of ODEs near periodic orbits can be studied via Poincaré maps, as sketched in §2.4.4. Maps (or mappings) can also be defined directly in modeling phenomena such as populations, which may more naturally be described in terms of generations, or via census data taken periodically. Dynamical systems can be *deterministic*, such as the ODEs introduced above and below, in which

the evolution rules specify uniquely the future and past, given the present state. Alternatively, in *stochastic* or random dynamical systems, an element of chance intervenes, often modeled by a gaussian (white noise) process. We shall meet simple examples of these in the form of stochastic differential equations (SDEs) in §6, and in particular, the drift-diffusion process.

2.4.1 Stable and unstable manifolds

In §2.1 we described the linearization procedure. We now investigate the *geometry* implied by the linear algebra. The eigenvectors of the linear system (2.1.4) (or (2.1.3)) define invariant subspaces⁷ of solutions described by the flow map (2.1.7) and exponential formulae such as (2.1.5-2.1.6). For convenience, we repeat (with a superscript e to denote that linearization is carried out at the equilibrium \mathbf{x}^e):

$$\mathbf{x}(t) = \Phi^e(t)\mathbf{x}_0, \text{ where } \Phi(t) = \mathbf{X}^e(t)[\mathbf{X}^e(0)]^{-1}, \quad (2.4.2)$$

where $\mathbf{X}^e(t)$ is a fundamental solution matrix for the linearised system (2.1.3).

It is a wonderful fact that the closed form solution (2.4.2) of the linearized system (2.1.4) is a good approximation for solutions of (2.0.1) as long as all the eigenvalues of $D\mathbf{f}(\mathbf{x}^e)$ have non-zero real parts and $|\xi|$ remains sufficiently small. More precisely, the Hartman–Grobman theorem [100, 89] asserts that there is a continuous change of coordinates transforming solutions of the full nonlinear system (2.0.1) in a neighborhood $B(\mathbf{x}^e)$ of the fixed point into those of (2.1.4) in a similar neighborhood of $\xi = \mathbf{0}$. This is called *topological equivalence*. Even better, the decomposition of the phase space \mathbb{R}^n into invariant subspaces spanned by (collections of) eigenvectors \mathbf{v}_i also holds for the nonlinear system in $B(\mathbf{x}^e)$, in that it possesses *invariant manifolds* filled with solutions whose qualitative behavior echoes that of the linearised system. (For our puposes, we need only know that a manifold is a curved space that looks locally like a piece of Euclidean space: see the examples below.)

First, recall that under the linear flow map $\Phi^e(t)$ (2.0.3) any solution initially lying in a k -dimensional subspace $\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, ($k < n$) remains in that subspace. Thus we can define two distinguished invariant subspaces whose names will reflect the asymptotic properties of solutions belonging to them. Suppose that $D\mathbf{f}(\mathbf{x}^e)$ has s ($\leq n$) and u ($\leq n - s$) eigenvalues with negative and positive real parts respectively and number them $\lambda_1, \dots, \lambda_s$ and $\lambda_{s+1}, \dots, \lambda_{s+u}$ (counting multiplicities). Let

$$E^s = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_s\} \text{ and } E^u = \text{span}\{\mathbf{v}_{s+1}, \dots, \mathbf{v}_{s+u}\}.$$

We call E^s the *stable subspace* and E^u the *unstable subspace*: solutions in E^s decay exponentially and those in E^u grow exponentially as t increases. The *stable manifold theorem* states that for (2.0.1) in a neighborhood $B(\mathbf{x}^e)$ of the equilibrium \mathbf{x}^e , there exist *local stable and unstable manifolds* $W_{loc}^s(\mathbf{x}^e)$, $W_{loc}^u(\mathbf{x}^e)$ of dimensions s and u respectively, tangent at \mathbf{x}^e to E^s and E^u , and characterised as follows, using the flow map (2.0.3):

$$\begin{aligned} W_{loc}^s(\mathbf{x}^e) &= \{\mathbf{x} \in B(\mathbf{x}^e) | \phi_t(\mathbf{x}) \rightarrow \mathbf{x}^e \text{ as } t \rightarrow +\infty \text{ and } \phi_t(\mathbf{x}) \in B(\mathbf{x}^e) \text{ for all } t \geq 0\}; \\ W_{loc}^u(\mathbf{x}^e) &= \{\mathbf{x} \in B(\mathbf{x}^e) | \phi_t(\mathbf{x}) \rightarrow \mathbf{x}^e \text{ as } t \rightarrow -\infty \text{ and } \phi_t(\mathbf{x}) \in B(\mathbf{x}^e) \text{ for all } t \leq 0\}. \end{aligned}$$

⁷A set \mathcal{S} is *invariant* for an ODE if every solution $\mathbf{x}(t)$ with initial condition $\mathbf{x}(0) \in \mathcal{S}$, remains in \mathcal{S} for all time t .

In words, the local stable manifold consists of all solutions that start and remain near the equilibrium for all future time and approach it as time tends to infinity, and the unstable manifold is defined similarly, with the substitutions “past time” and “minus infinity.”

These manifolds are smooth, curved surfaces, which locally look like the linear subspaces E^s and E^u (as the surface of the earth looks locally like a plane); see Fig. 25(a). We can make this more precise as follows. In terms of the local coordinates ξ near \mathbf{x}^e the smooth manifolds $W_{loc}^s(\mathbf{x}^e)$ and $W_{loc}^u(\mathbf{x}^e)$ can be expressed as graphs over E^s and E^u respectively. Picking an eigenvector basis and letting $E^{s\perp}$ denote the $(n-s)$ -dimensional orthogonal complement to E^s and $\mathbf{y} \in E^s$, $\mathbf{z} \in E^{s\perp}$ be local coordinates, we can write

$$W_{loc}^s(\mathbf{x}^e) = \{(\mathbf{y}, \mathbf{z}) | (\mathbf{y}, \mathbf{z}) \in B(\mathbf{0}) \text{ and } \mathbf{z} = \mathbf{g}(\mathbf{y})\} \quad (2.4.3)$$

for some smooth function $\mathbf{g} : E^s \rightarrow E^{s\perp}$. We cannot generally compute \mathbf{g} (if we could we would have found solutions of (2.0.1)), but we can approximate it, as we shall see below.

Equipped with the local manifolds we can define the *global* stable and unstable manifolds:

$$W^s(\mathbf{x}^e) = \bigcup_{t \leq 0} \phi_t(W_{loc}^s(\mathbf{x}^e)) ; \quad W^u(\mathbf{x}^e) = \bigcup_{t \geq 0} \phi_t(W_{loc}^u(\mathbf{x}^e)) .$$

These are the unions of backwards and forwards images of the local manifolds under the (nonlinear) flow map. Thus $W^s(\mathbf{x}^e)$ is the set of *all* points whose orbits approach \mathbf{x}^e as $t \rightarrow +\infty$, even if they leave $B(\mathbf{x}^e)$ for a while, and $W^u(\mathbf{x}^e)$ is the set of all points whose orbits approach \mathbf{x}^e as $t \rightarrow -\infty$. Stable manifolds cannot intersect themselves (except at the fixed points) or other stable manifolds, since this would violate uniqueness of solutions (the intersection point would lead to more than one future). The same is true of unstable manifolds. However, intersections of stable and unstable manifolds do occur: they lie on solutions that lead from one fixed point to another. Self-intersection points are called *homoclinic*: $p \in W^u(\mathbf{x}^e) \cap W^s(\mathbf{x}^e)$; when manifolds belonging to different points intersect, they are called *heteroclinic*: $p \in W^u(\mathbf{x}^{e_1}) \cap W^s(\mathbf{x}^{e_2})$. See Fig. 25.

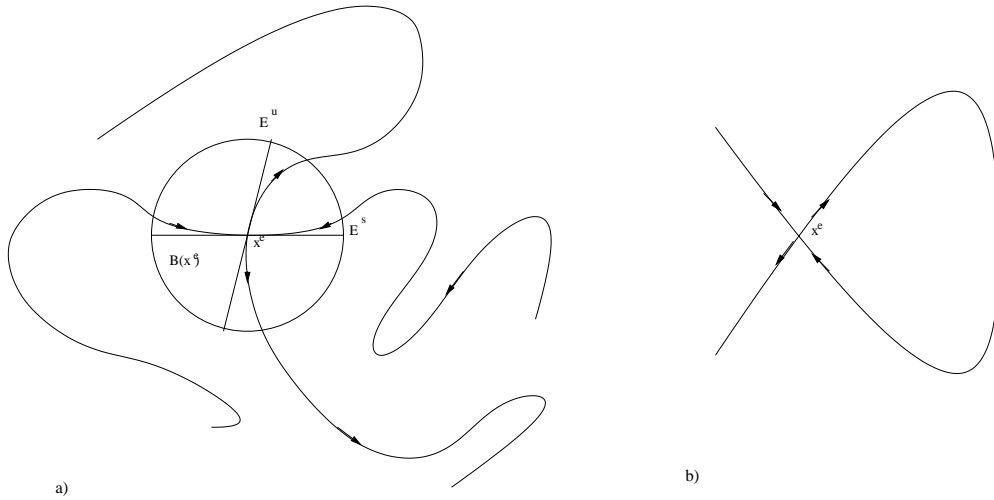


Figure 25: Stable and unstable manifolds (a). The local manifolds in $B(\mathbf{x}^e)$ can be expressed as graphs as in (2.4.3), but globally they may “double back” and even intersect one another. A homoclinic orbit (b).

2.4.2 Planar systems of ODEs

In a similar manner to fixed points, periodic orbits also carry stable and unstable manifolds filled with orbits that are forwards and backwards-asymptotic to them. The global orbit structure in state space is largely organized by the stable and unstable manifolds. The former can act as separatrices dividing solutions that can have very different future behaviors. See example 7 below. However, while Taylor series approximations for graphs defining the local manifolds (such as (2.4.3)) can be found, truly global behavior can only be established in simple cases such as two-dimensional systems, of the form

$$\begin{aligned}\dot{x}_1 &= f_1(x_1, x_2), \\ \dot{x}_2 &= f_2(x_1, x_2).\end{aligned}\tag{2.4.4}$$

Nullclines are the curves on which one component of the vectorfield vanishes so that solutions cross them either vertically or horizontally. Specifically, $\dot{x}_1 = 0$ on $f_1(x_1, x_2) = 0$ and $\dot{x}_2 = 0$ on $f_2(x_1, x_2) = 0$. (One can also define *isoclines* [231] on which the slope of the vectorfield is constant: the nullclines are simply a special case of these.) It is often possible to solve for the nullclines explicitly, plot them on the state (or phase-) plane \mathbb{R}^2 , and then sketch enough arcs and segments of solution curves that one can better interpret the results of numerical solutions. Clearly, fixed points lie at the intersections of the nullclines, where both components of the vectorfield vanish, and the orientations of the tangent vectors in regions bounded by different components of the nullclines can help us to assemble the local pieces of the state space determined by linearization into a global picture.

Example 6. *Nonlinear gain control in the retina [231, §6.2].*

In this example we illustrate linearization and the use of nullclines, and show how some global implications can be deduced from simple calculations of the direction of the vectorfield. Here the state variables B and A represent activity levels of bipolar and amacrine cells in the retina, exposed to light of intensity $L \geq 0$ (assumed to be constant):

$$\begin{aligned}\dot{B} &= \frac{1}{\tau_B}[-B + L/(1 + A)] \\ \dot{A} &= \frac{1}{\tau_A}[-A + 2B]\end{aligned}\tag{2.4.5}$$

The terms $-A, -B$ model the decay of activity with time constants τ_A and τ_B , in the absence of inputs, $2B$ models the excitatory influence of bipolar cells on amacrine cells, and $L/(1 + A)$ in the first equation represents the feedback from amacrine cells that reduces the bipolar response. We shall assume that L is constant. See [231] for more details.

Neural activity, measured as an average spike rate for example, is normally regarded as a non-negative quantity (it might drop to zero), so we first ask if this model allows A or B to become negative. Suppose we start at a point $(A, B) = (A(0), 0)$ with $A(0) \geq 0$: then $\dot{B} = \frac{L}{(\tau_B(1+A))} \geq 0$, so B cannot go negative. Similarly, starting at $(0, B(0))$ with $B(0) \geq 0$, $\dot{A} = \frac{2B}{\tau_A} \geq 0$ and A cannot go negative either. In fact the positive quadrant $\mathcal{Q}^+ = \{(B, A) | B \geq 0, A \geq 0\}$ is positively invariant: solutions of (2.4.5) starting in \mathcal{Q}^+ stay in \mathcal{Q}^+ for all future time. This also ensures that, given physically acceptable initial conditions $(A(0) \geq 0, B(0) \geq 0)$, solutions stay away from the line $A = -1$ where the function defining the first equation is singular (= blows up). See Fig. 26.

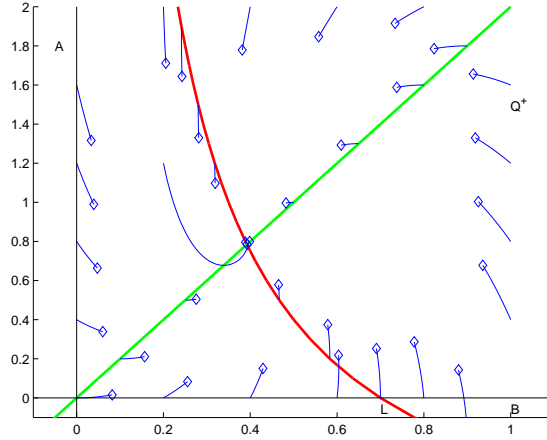


Figure 26: The phase portrait of the retinal gain model. The red hyperbola and green diagonal are the nullclines $\dot{A} = 0 \Rightarrow B = L/(1 + A)$ and $\dot{B} = 0 \Rightarrow A = 2B$ respectively.

Fixed points are given by $-A + 2B = 0$ and $-B + L/(1 + A) = 0$, and setting $A = 2B$ in the second equation and solving for B we obtain:

$$\bar{B} = \frac{-1 \pm \sqrt{1 + 8L}}{4} \quad \text{and} \quad \bar{A} = \frac{-1 \pm \sqrt{1 + 8L}}{2}, \quad (2.4.6)$$

so there is a single equilibrium in the physically-relevant region \mathcal{Q}^+ , corresponding to choosing the positive roots in (2.4.6). Linearizing (2.4.5) we obtain the matrix:

$$\begin{pmatrix} \frac{-1}{\tau_B} & \frac{-L}{\tau_B(1+A)^2} \\ \frac{2}{\tau_A} & \frac{-1}{\tau_A} \end{pmatrix}, \quad (2.4.7)$$

and evaluating at (\bar{B}, \bar{A}) yields:

$$\begin{pmatrix} \frac{-1}{\tau_B} & \frac{(\sqrt{1+8L}-1)^2}{16\tau_B L} \\ \frac{2}{\tau_A} & \frac{-1}{\tau_A} \end{pmatrix} \Rightarrow \det(D\mathbf{f}) = \frac{1 + 8L - \sqrt{1 + 8L}}{4\tau_A\tau_B L} > 0, \quad \text{tr}(D\mathbf{f}) = -\left(\frac{1}{\tau_A} + \frac{1}{\tau_B}\right) < 0. \quad (2.4.8)$$

The eigenvalues of the 2×2 matrix $D\mathbf{f}$ are given by the quadratic equation

$$\lambda^2 - \text{tr}(D\mathbf{f})\lambda + \det(D\mathbf{f}) = 0, \quad (2.4.9)$$

and so, for all $L, \tau_A, \tau_B > 0$, we conclude that the fixed point is stable. (See Fig. 27 and the discussion on pp 42-23 below.)

We have already shown that the positive quadrant \mathcal{Q}^+ is positively invariant. We can also show that a trapping region \mathcal{A} exists within \mathcal{Q}^+ such that all solutions starting in \mathcal{Q}^+ sufficiently close to \mathcal{A} eventually enter \mathcal{A} . Referring to Fig. 26, we draw vertical and horizontal lines at $B = C$ and $A = 2C$ for some $C > L$ that together with the A and B axes bound a rectangular domain \mathcal{A} and consider the components of the vector field on these lines and normal to them:

$$\begin{aligned} \dot{B} |_{B=C} &= \frac{1}{\tau_B} [-C + L/(1 + A)] < 0, & \text{since } A \geq 0 \text{ and } C > L; \\ \dot{A} |_{A=2C} &= \frac{1}{\tau_A} [-2C + 2B] < 0, & \text{since } B \leq C. \end{aligned}$$

With the results established earlier, this shows that all solutions starting on the boundary of \mathcal{A} enter it. In fact we can do even more, but first we need another result for two-dimensional systems.

Theorem 2 (Bendixson's criterion). *If on a simply-connected region $D \subset \mathbb{R}^2$ the quantity $\text{tr}(D\mathbf{f}) = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}$ is not identically zero and does not change sign, then the planar ODE (2.4.4) has no closed orbits lying entirely in D .*

Proof. (Sketch) This is proved by noting that on any closed (= time-periodic) orbit Γ we have:

$$\frac{\dot{x}_2}{\dot{x}_1} = \frac{dx_2}{dx_1} = \frac{f_2}{f_1} \Rightarrow \int_{\Gamma} [f_1(x_1, x_2) dx_2 - f_2(x_1, x_2) dx_1] = 0;$$

hence, appealing to Green's theorem, we deduce that

$$\int \int_{\text{int}\Gamma} \left[\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} \right] dx_1 dx_2 = 0,$$

thereby obtaining a contradiction. Simple connectivity is required for Green's theorem. □

Exercise 2. *Use Bendixson's criterion to show that the ODEs (2.4.5) cannot have periodic solutions. Can you go on from this to prove that all solutions starting in the positive quadrant approach the stable sink (\bar{B}, \bar{A}) ?*

Note: In the first version of these notes PH thought he had a solution to the following exercise:

Exercise 2.0. Construct a Liapunov function and thereby prove that the fixed point (\bar{B}, \bar{A}) is globally asymptotically stable. Hint: Let $B = \bar{B} + x$, $A = \bar{A} + y$, transforming the fixed point to $(x, y) = (0, 0)$, consider the quadratic form $V(x, y) = x^2 + 2\beta xy + \gamma y^2$, and try to choose β and γ appropriately.

He was mistaken. The ODEs in the new variables are

$$\dot{x} = \frac{1}{\tau_B} \left[\frac{-4(a+1)x - 2(a-1)y - 8xy}{4(a+1+2y)} \right], \quad \dot{y} = \frac{1}{\tau_A} [2x - y],$$

where $a = \sqrt{1+8L}$. The denominator in the first ODE is strictly positive in the positive quadrant of the (B, A) -plane, which simplifies finding the sign of \dot{V} , but the quadratic term in the numerator leads to cubic terms in the expression for \dot{V} , which precludes negative definiteness. Perhaps one can find a higher order polynomial function that works but it'll be tricky! See Exercise 5 below for a Liapunov function that does work.

The extended Definition 5 of stability given at the end of §2.2 suggests the more general idea of an *attracting set*, which subsumes the special cases of asymptotically stable fixed points (sinks) and stable periodic orbits or limit cycles:

Definition 6. *An invariant set \mathcal{A} for a dynamical system is an **attracting set** if all solutions starting in a neighborhood of \mathcal{A} , approach \mathcal{A} as $t \rightarrow \infty$.*

The Poincaré-Bendixson theorem, stated at the beginning of §2.4.4, provides a kind of converse to Bendixson's criterion in that it gives conditions sufficient for the existence of closed orbits.

2.4.3 Center manifolds and local bifurcations

As parameters change, so do phase portraits and the resulting dynamical behavior of model systems. New fixed points can appear, and the stability types of existing ones may change in *bifurcations*. Examples will follow (the reader may like to look ahead to Example 7 and Figs. 33-34 at the end of this section), but first we develop some background. Bifurcation theory addresses how system behavior changes as *control parameters* are varied. We consider *families* of systems of the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}; \boldsymbol{\mu}) \quad \text{or} \quad \dot{\mathbf{x}} = \mathbf{f}_{\boldsymbol{\mu}}(\mathbf{x}) , \quad (2.4.10)$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ is a (vector of) parameter(s). Often all but one are (temporarily) fixed and we consider a one-parameter family, which corresponds to a path in the appropriate space of systems.

Here it is necessary to introduce the important notion of *structural stability*, which is entirely different from the Liapunov and asymptotic stabilities of equilibria and orbits defined in §2.1-2.2. Rather than considering how solutions of a given ODE system change under small perturbations of the initial conditions, we ask how the entire set of solutions – the global phase portrait – of a system changes, when the defining vectorfield $\mathbf{f}_{\boldsymbol{\mu}}(\mathbf{x})$ undergoes a small change, as happens when parameters vary. We will sweep the (tricky) technical issues of defining “small perturbation” for a system of ODEs under the rug (one needs to define a space of systems with appropriate metrics and norms to describe how close two systems are, etc.), and merely sketch the idea. Rigorous definitions can be found in [89].

We say that two systems are *topologically equivalent* if the phase portrait of one can be mapped onto that of the other by a continuous coordinate change: i.e., there is a one-to-one correspondence of all orbit segments. Furthermore, the sense of time must be preserved (sinks remain sinks, sources remain sources). A *small perturbation* is one for which not only function values, but also their first derivatives, are close:

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{g}(\mathbf{x})\| < \epsilon \quad \text{and} \quad \|D\mathbf{f}(\mathbf{x}) - D\mathbf{g}(\mathbf{x})\| < \epsilon. \quad (2.4.11)$$

Thus equipped, we may state:

Definition 7. *The nonlinear system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is **structurally stable** if all sufficiently close systems $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ are topologically equivalent to $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$.*

The case of two-dimensional linear systems nicely illustrates this, and further illuminates our linearized stability analyses. Consider the ODEs

$$\begin{aligned} \dot{x}_1 &= a_{11}x_1 + a_{12}x_2 \\ \dot{x}_2 &= a_{21}x_1 + a_{22}x_2 . \end{aligned} \quad (2.4.12)$$

The eigenvalues of the matrix $\mathbf{A} = [a_{ij}]$ are given by the roots of the characteristic equation

$$\lambda^2 - \lambda \operatorname{tr} \mathbf{A} + \det \mathbf{A} = 0 ,$$

where $\operatorname{tr} \mathbf{A} = a_{11} + a_{22}$ and $\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$. Thus, if $\det(\mathbf{A}) < 0$ then $(0,0)$ is a saddle point, while if $\det(\mathbf{A}) > 0$ and $\operatorname{tr}(\mathbf{A}) < 0$ it is a sink and if $\det(\mathbf{A}) > 0$ and $\operatorname{tr}(\mathbf{A}) > 0$ it is a source. The diagram of Fig. 27 summarises the qualitative behavior. On the diagram we also indicate the parabola $[\operatorname{tr}(\mathbf{A})]^2 = 4 \det(\mathbf{A})$ on which the eigenvalues change from real to complex, but this

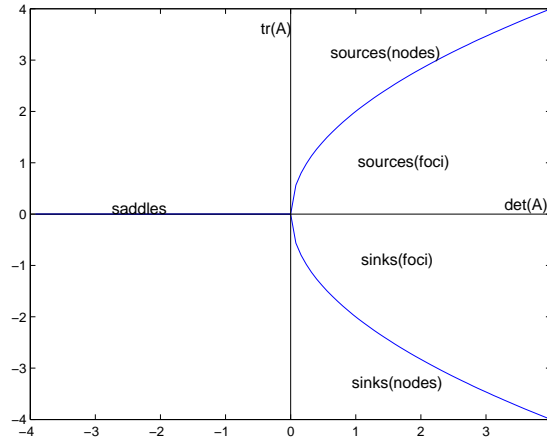


Figure 27: Stability types of the fixed point at the origin for the planar system (2.4.12).

does not correspond to a *topological* change in the character of solutions: a sink is a sink, whether solutions approach it directly or in spirals.

Here the space of systems is effectively the plane with $\det(\mathbf{A})$, $\text{tr}(\mathbf{A})$ as coordinates and any system not on the $\text{tr}(\mathbf{A})$ -axis or the positive $\det(\mathbf{A})$ -axis is structurally stable, since all systems in a sufficiently small neighborhood of it share its behavior. On the positive $\det(\mathbf{A})$ -axis, however, we have purely imaginary eigenvalues and the ODE behaves like an undamped harmonic oscillator. Such systems are structurally unstable. Moving off the axis – adding damping, no matter how small – yields qualitatively different behavior. The periodic orbits are all destroyed and replaced by solutions spiralling into a sink, or out from a source.

We now return to bifurcations, via a simple example. The one-parameter, one-dimensional family of ODEs

$$\dot{x} = \mu - x^2 \quad (2.4.13)$$

will serve to introduce local bifurcation theory. Equation (2.4.13) is soluble in closed form, but here we need only observe that, for $\mu < 0$ it has no equilibria and all solutions approach $-\infty$ as t increases and $+\infty$ as t decreases, while for $\mu > 0$ there are two hyperbolic equilibria $x_{e1} = +\sqrt{\mu}$ and $x_{e2} = -\sqrt{\mu}$, the former being an asymptotically stable sink and the latter an unstable source, and that all solutions starting at $x(0) > -\sqrt{\mu}$ approach the sink as $t \rightarrow +\infty$. At $\mu = 0$ the single equilibrium $x_e = 0$ is non-hyperbolic and the linearisation $\dot{\xi} = 2x_e\xi = 0$ gives no information on stability. For all $\mu \neq 0$ (2.4.13) is, in fact, structurally stable: its behavior changes as it passes through the point $\mu = 0$ of structural instability. This information can be conveniently pictured on a *bifurcation diagram*: a plot of the branches of equilibria versus the parameter: Fig. 28. This is a *saddle-node* bifurcation [100, 89].

In general we have:

Definition 8. A value μ_0 of μ for which Equation (2.4.10) is structurally unstable is a bifurcation value.

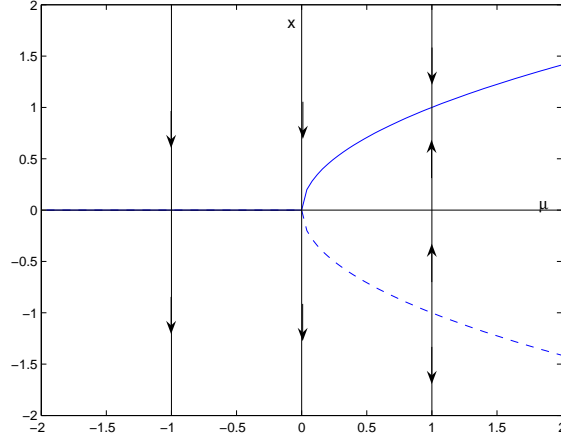


Figure 28: The bifurcation diagram for Equation (2.4.13). Vertical lines represent phase portraits of individual systems, with arrows corresponding to time and thus displaying stability; — represents a branch of sinks, - - - a branch of sources.

The bifurcation at $\mu = 0$ for (2.4.13) is called *local* because for $\mu \approx 0$ the qualitative change in the phase portrait is confined to a neighborhood of $x = 0$. In fact when such bifurcations occur in more complex, higher-dimensional systems, they can be analysed by local methods: examination of truncated Taylor series expansions near the degenerate (non-hyperbolic) equilibrium or periodic orbit involved in the bifurcation. Here the *center subspace* E^c , spanned by eigenvectors whose eigenvalues have zero real part, is important. For example, supplementing the example of (2.4.13) in the following way:

$$\begin{aligned}\dot{x}_1 &= -x_1 \\ \dot{x}_2 &= \mu - x_2^2, \\ \dot{x}_3 &= 2x_3\end{aligned}\tag{2.4.14}$$

we can see that “nothing happens” in the x_1 - and x_3 -directions as μ varies. At $\mu = 0$ the degenerate equilibrium $\mathbf{x} = \mathbf{0}$ is hyperbolic in those two directions and they can effectively be ignored. It is a remarkable fact that this remains true even when there is nonlinear coupling between the hyperbolic and non-hyperbolic directions. This is the content of the *center manifold theorem* [89]. In addition to local stable and unstable manifolds $W_{loc}^s(\mathbf{x}^e)$, $W_{loc}^u(\mathbf{x}^e)$, a center manifold $W_{loc}^c(\mathbf{x}^e)$, tangent to E^c , exists in a neighborhood of the degenerate equilibrium \mathbf{x}^e . See Fig. 29. Unlike $W_{loc}^s(\mathbf{x}^e)$ and $W_{loc}^u(\mathbf{x}^e)$, however, $W_{loc}^c(\mathbf{x}^e)$ cannot be characterized in terms of the asymptotic behavior of solutions within it. For example, at $\mu = 0$ solutions of (2.4.14) approach $\mathbf{x} = \mathbf{0}$ from $x_2 < 0$ and recede from it for $x_2 > 0$ (the fixed point is “semi-stable”).

Fig. 29 provides a schematic picture of the phase space near a degenerate equilibrium with s negative, u positive, and c zero (real part) eigenvalues. As the picture suggests, the “nonlinear coordinates” implicit in the definitions of W_{loc}^s , W_{loc}^u , and W_{loc}^c allow us to separate stable, unstable and non-hyperbolic or bifurcation behaviors and so to *reduce* our analysis to that of a c -dimensional system restricted to the center manifold.

To describe the reduction process we assume that coordinates have been chosen with the degenerate equilibrium \mathbf{x}_e at the origin and such that the matrix $\mathbf{Df}(\mathbf{x}_e)$ is block diagonalised. Thus

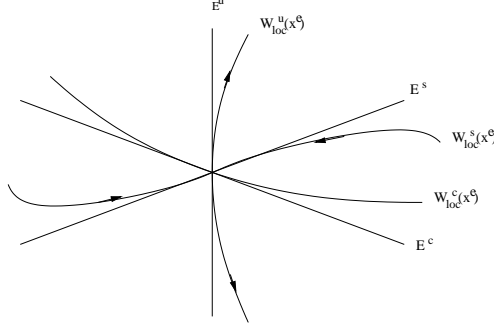


Figure 29: The stable, center, and unstable manifolds

(2.4.10) can be written in the form:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{y}), \\ \dot{\mathbf{y}} &= \mathbf{B}\mathbf{y} + \mathbf{g}(\mathbf{x}, \mathbf{y}),\end{aligned}\tag{2.4.15}$$

where \mathbf{x} belongs to the center subspace E^c and \mathbf{y} belongs to the stable and unstable subspaces. We drop the explicit reference to parameter dependence. Here all the eigenvalues of the $c \times c$ matrix \mathbf{A} have zero real parts and the $(n - c) \times (n - c) = (s + u) \times (s + u)$ matrix \mathbf{B} has only eigenvalues with non-zero real parts. The center manifold can now be expressed as a graph \mathbf{h} over E^c :

$$\mathbf{y} = \mathbf{h}(\mathbf{x}).\tag{2.4.16}$$

This implies that, as long as solutions remain on the center manifold, the full state (\mathbf{x}, \mathbf{y}) of the system can be specified, via (2.4.16) by the state of the \mathbf{x} variables alone. The *reduced system* is then defined as

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{h}(\mathbf{x}))\tag{2.4.17}$$

and the stability and bifurcation behavior near the degenerate equilibrium can be deduced from it.

As an example, consider the two dimensional system

$$\begin{aligned}\dot{x}_1 &= \mu x_1 - x_1 x_2, \\ \dot{x}_2 &= -x_2 + x_1^2,\end{aligned}\tag{2.4.18}$$

whose linear part is already diagonalised. At $\mu = 0$ the eigenvalues are 0 and -1 respectively and $E^c = \text{span}\{(1, 0)\}$, $E^s = \text{span}\{(0, 1)\}$, so we seek the center manifold as a function $x_2 = h(x_1)$. This function will be a particular solution of the ODE

$$\frac{dx_2}{dx_1} = \frac{-x_2 + x_1^2}{-x_1 x_2} = \frac{1}{x_1} - \frac{x_1}{x_2},\tag{2.4.19}$$

which cannot be solved explicitly. Here and in general we instead approximate the graph $h(x_1)$ as a Taylor series, which, in view of the local nature of our study, will be adequate provided that the leading terms do not vanish.

Returning to the general problem (2.4.15), we substitute $\mathbf{y} = \mathbf{h}(\mathbf{x})$ into the second component and use the chain rule and the first component to obtain a partial differential equation for the function $\mathbf{h}(\mathbf{x})$:

$$D\mathbf{h}(\mathbf{x})[\mathbf{A}\mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{h}(\mathbf{x}))] = \mathbf{B}\mathbf{h}(\mathbf{x}) + \mathbf{g}(\mathbf{x}, \mathbf{h}(\mathbf{x}))\tag{2.4.20}$$

with “boundary conditions”

$$\mathbf{h}(\mathbf{0}) = \mathbf{0} ; \quad D\mathbf{h}(\mathbf{0}) = \mathbf{0} , \quad (2.4.21)$$

which result from the tangency of W_{loc}^c to E^c at $\mathbf{x}_e = \mathbf{0}$. We shall seek an approximate solution of (2.4.20) and (2.4.21) in the form of a Taylor series. For our example (2.4.18) at $\mu = 0$, therefore, we set

$$\begin{aligned} x_2 = h(x_1) &= a_2 x_1^2 + a_3 x_1^3 + \mathcal{O}(|x_1|^4) \\ \Rightarrow Dh(x_1) &= 2a_2 x_1 + 3a_3 x_1^2 + \mathcal{O}(|x_1|^3) ; \end{aligned} \quad (2.4.22)$$

due to the boundary conditions, the Taylor series necessarily starts at second order. Equation (2.4.20) becomes, in this case,

$$\begin{aligned} [2a_2 x_1 + 3a_3 x_1^2 + \mathcal{O}(|x_1|^3)] \{-x_1 [a_2 x_1^2 + a_3 x_1^3 + \mathcal{O}(|x_1|^4)]\} \\ = -[a_2 x_1^2 + a_3 x_1^3 + \mathcal{O}(|x_1|^4)] + x_1^2 . \end{aligned} \quad (2.4.23)$$

Equating terms of comparable orders in $|x_1|$ yields:

$$\left. \begin{aligned} \mathcal{O}(|x_1|^2) : 0 &= -a_2 + 1 \\ \mathcal{O}(|x_1|^3) : 0 &= a_3 \end{aligned} \right\} \quad (2.4.24)$$

and we conclude that the function h may be written

$$x_2 = h(x_1) = x_1^2 + \mathcal{O}(|x_1|^4) . \quad (2.4.25)$$

At $\mu = 0$ the reduced equation is therefore

$$\dot{x}_1 = -x_1 [x_1^2 + \mathcal{O}(|x_1|^4)] = -x_1^3 + \mathcal{O}(|x_1|^5) , \quad (2.4.26)$$

the behavior of which is dominated, for small $|x_1|$, by the negative cubic term which pushes solutions towards $x_1 = 0$. The degenerate equilibrium is evidently stable at the bifurcation point, as well as for all $\mu < 0$. This is not directly obvious from (2.4.18): the linearisation $\dot{x}_1 = \mu x_1$ of the first component tells us nothing at $\mu = 0$ and the reader can check, repeating the calculations above, that a change in sign in the term x_1^2 in the second component turns stability on its head! (See Example 1 above.)

To study the bifurcation that occurs at as $(x_1, x_2) = (0, 0)$ for (2.4.18) as μ varies from zero, we should really incorporate μ as an additional “dummy state variable” and compute an extended center manifold tangent to the (x_1, μ) -plane, but in this case we can solve explicitly for the new equilibria that appear for $\mu > 0$. Representative phase portraits of the full system for $\mu < 0$, $\mu = 0$ and $\mu > 0$ appear in Fig. 30, and Fig. 31 shows the bifurcation diagram for the reduced system. This is an example of a *pitchfork* bifurcation [89].

Exercise 3. *Verify by direct calculation of equilibria, and by checking stability types, that the phase portraits of Fig. 30 and the bifurcation diagram Fig. 31 are correct.*

Example 7. *The Usher-McClelland leaky accumulator model for two-alternative decisions [214].*

This example illustrates nullclines, invariant manifolds and local bifurcations. It is similar to the short-term memory circuit of [231, §6.4] and the “winner take all” decision-making model of Eqn. (2.5.2) below. Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_1 - f(x_2; g, \beta) + s_1, \\ \dot{x}_2 &= -x_2 - f(x_1; g, \beta) + s_2, \end{aligned} \quad (2.4.27)$$

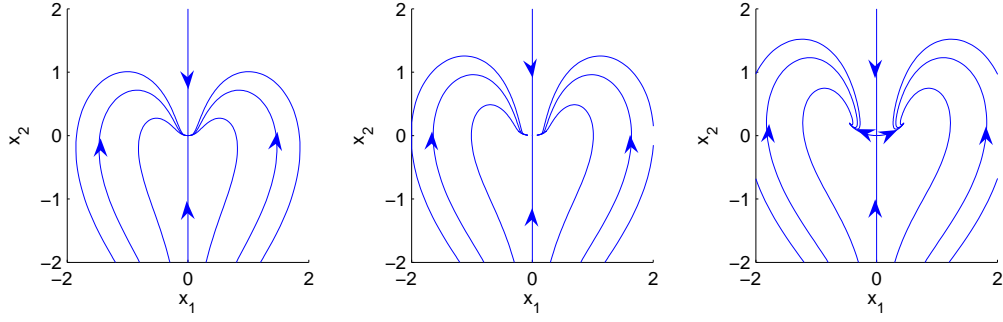


Figure 30: Phase portraits of the two dimensional system (2.4.18): (a) $\mu = -0.2$, (b) $\mu = 0$, (c) $\mu = 0.2$.

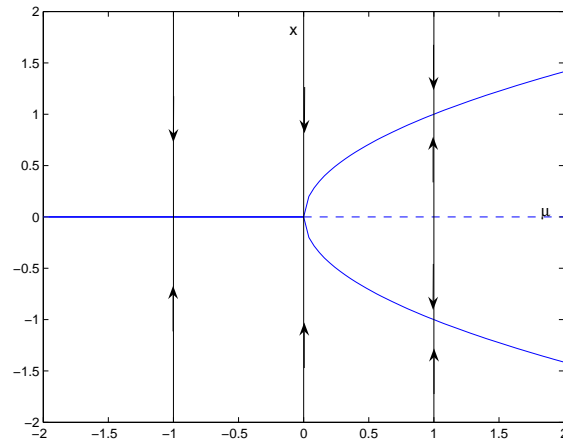


Figure 31: The bifurcation diagram for Equation (2.4.18); — represents a branch of sinks, - - - a branch of saddle points.

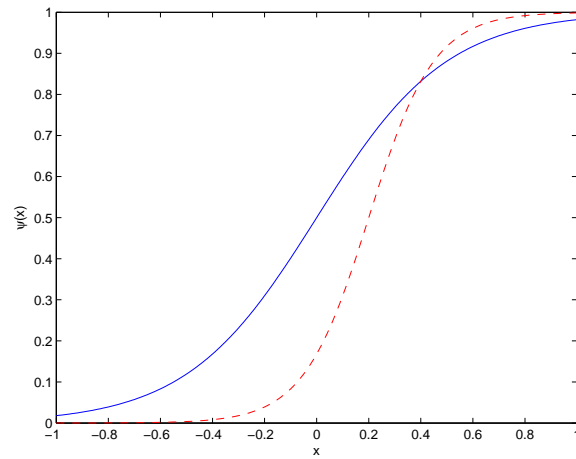


Figure 32: Sigmoidal or logistic activation functions, showing the effects of gain g and bias β . Solid blue curve: $g = 1$, $\beta = 0$; dashed red curve: $g = 2$, $\beta = 0.2$.

where the function f is a sigmoid:

$$f(x) = \frac{1}{1 + \exp(-4g(x - \beta))}; \quad (2.4.28)$$

see Fig. 32. Here x_1 and x_2 represent the activity levels of two pools of neurons that mutually inhibit each other through the “input/output” function f . The neurons in each pool are selectively responsive to one of the stimuli s_j , as, for example, in oculo-motor brain areas of monkeys trained to discriminate arrays of dots moving to the right or left [93, 194]. In the absence of stimuli ($s_j = 0$), the activities $x_j(t)$ decay to a baseline level due to the leakage terms $-x_j$. With stimuli of different magnitudes present (e.g. $s_2 > s_1$) one unit receives greater input than the other, so that its activity initially grows faster than the other’s; this effect is magnified by the greater inhibition that it exerts (if $x_2 > x_1$ then $f(x_2) > f(x_1)$). The decision is made either when x_1 or x_2 first crosses a threshold, or at a specific time, by selecting the pool with higher activity. See §6.1 for more on models of this type.

Eqn. (2.4.27) has four parameters: g , β , s_1 and s_2 : g and β control the gain and offset or bias implicit in the (synaptic) connections between the two pools and they can be expected to change only slowly (as the animal learns the task); the stimulus levels s_j will change from trial to trial. The nullclines of (2.4.27) can be read off the right hand sides immediately as explicit functions of x_1 in terms of x_2 and vice versa:

$$\dot{x}_1 = 0 \text{ on } x_1 = s_1 - f(x_2; g, \beta) \text{ and } \dot{x}_2 = 0 \text{ on } x_2 = s_2 - f(x_1; g, \beta), \quad (2.4.29)$$

and, depending upon the parameter values, we can produce pictures such as those of Fig. 33. Note that the system may have one, two or three fixed points. In particular, if gain g is sufficiently low then the slopes of the sigmoids are such that only one intersection is possible. For high gain and with sufficiently similar stimuli, three fixed points exist.

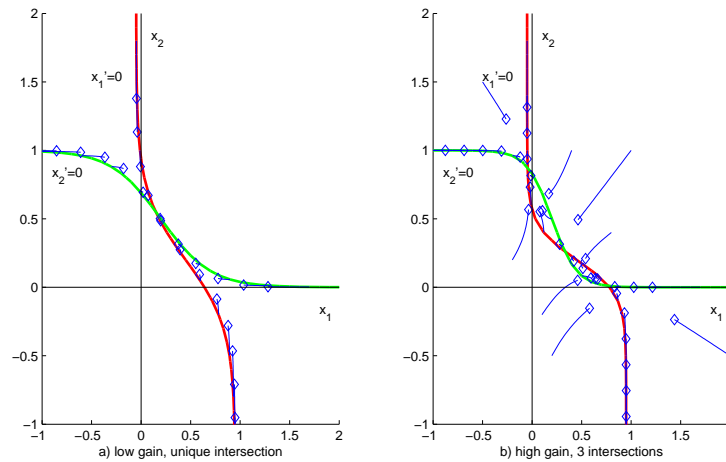


Figure 33: Examples of nullclines and fixed points for equations (2.4.27) with $s_1 = 0.95$, $s_2 = 1$, $\beta = 0.2$, and $g = 1$ (on left) and $g = 2$ (on right).

Linearizing (2.4.27) we obtain the matrix

$$D\mathbf{f} = \begin{pmatrix} -1 & -f'(x_2; g, \beta) \\ -f'(x_1; g, \beta) & -1 \end{pmatrix}, \quad (2.4.30)$$

where the derivatives $f' = f'(\bar{x}_j; g, \beta)$ are evaluated at the fixed point in question. From (2.4.29) we see that $\text{tr}(D\mathbf{f}) = -2$ and $\det(D\mathbf{f}) = 1 - f'(\bar{x}_1; g, \beta)f'(\bar{x}_2; g, \beta)$. Using the fact that the eigenvalues of a 2×2 matrix such as (2.4.30) are the roots of the quadratic equation

$$\begin{aligned} \lambda^2 - \text{tr}(D\mathbf{f})\lambda + \det(D\mathbf{f}) &= 0 \Rightarrow \lambda_{1,2} = \frac{\text{tr}(D\mathbf{f}) \pm \sqrt{[\text{tr}(D\mathbf{f})]^2 - 4\det(D\mathbf{f})}}{2} \\ &= -1 \pm \sqrt{f'(\bar{x}_1; g, \beta)f'(\bar{x}_2; g, \beta)}, \end{aligned} \quad (2.4.31)$$

we can deduce that the fixed points are saddles with one positive and one negative eigenvalue if $f'(\bar{x}_1; g, \beta)f'(\bar{x}_2; g, \beta) > 1$ and sinks, with two negative eigenvalues if $f'(\bar{x}_1; g, \beta)f'(\bar{x}_2; g, \beta) < 1$. The fixed points (\bar{x}_1, \bar{x}_2) must be found numerically by simultaneous solution of the transcendental equations (2.4.29) defining the nullclines, but it is easy to deduce stability from the graphical evidence of Fig. 34. For small slopes and a unique intersection of nullclines, we have $f'(\bar{x}_1; g, \beta)f'(\bar{x}_2; g, \beta) < 1$ and stability, while for large slopes an unstable fixed point can appear. The upshot is that, when $g > 1$ and multiple equilibria exist, the “outer” fixed points that lie in regions where the derivatives $f'(\bar{x}_j; g, \beta)$ are small, are sinks and the “inner” one is a saddle. See Fig. 34.

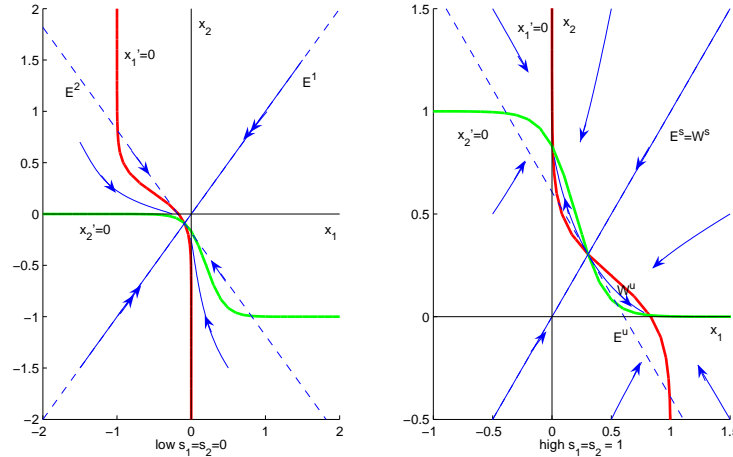


Figure 34: The phase planes of (2.4.27) for large gain $g = 2$, and equal stimuli, showing one stable fixed point in the absence of stimuli ($s_1 = s_2 = 0$ on left), and three for $s_1 = s_2 = 1$ (on right). In the latter case the outer fixed points are sinks and the inner one is a saddle.

In applying Usher and McClelland’s model to discriminating between two stimuli, this phase portrait may be interpreted as follows. In the absence of stimuli, or with low stimulus levels, both pools of neurons relax to baseline activity. When a stronger stimulus with low coherence is presented, so that no preponderant fraction of dots is moving to left or to right, both neural pools are equally activated, and the one that “wins” by inhibiting the other is determined by initial conditions. Informally, the animal chooses left or right “at random.”

Exercise 4. Show that, in the case of equal stimulus strengths $s_1 = s_2$, the phase portrait of Eqn. (2.4.27) is reflection-symmetric about the diagonal $x_1 = x_2$. What happens as the difference $s_1 - s_2$ changes between 0 and 1, if the sum $s_1 + s_2 = 1$ is fixed and $g > 1$? Sketch a bifurcation diagram and provide an interpretation in terms of typical behavior(s) as $t \rightarrow \infty$ (i.e., long-term decision outcomes).

Exercise 5. Show that the function

$$V(x_1, x_2) = \int^{x_1} x \frac{\partial f}{\partial x} dx + \int^{x_2} x \frac{\partial f}{\partial x} dx + f(x_1)f(x_2) \quad (2.4.32)$$

is a Liapunov function for Eqns. (2.4.27) with $s_1 = s_2 = 0$, in the sense that $dV/dt < 0$ unless $x_1 = -f(x_2)$ and $x_2 = -f(x_1)$ (here we suppress the explicit dependence of f on the parameters g and β). What does the function $V(x_1, x_2)$ look like? What can you deduce about the global structure of the state space? Can you generalize V to cover the case $s_1, s_2 \neq 0$?

Grossberg and Hopfield [45, 111, 112, 86] have shown that rather large classes of n -dimensional neural networks with symmetric connectivity matrices, including that of Example 7, admit Liapunov functions: see [231, §14.5, Theorem 16] (Exercise 5 is a very special two-dimensional case). This implies that all solutions run “downhill” and approach equilibria as time progresses, and for typical systems almost all solutions approach stable equilibria. No sustained periodic or other cyclic activity is possible. In contrast, consider the following generalization of Example 1 in §2.1 above:

Example 8. Consider the two-dimensional system

$$\begin{aligned} \dot{x} &= \mu x - \omega y + \alpha(x^2 + y^2)x, \\ \dot{y} &= \omega x + \mu y + \alpha(x^2 + y^2)y, \end{aligned} \quad (2.4.33)$$

where α and ω are regarded as fixed parameters and μ as a variable parameter. Passing to polar coordinates as in Example 1, we have:

$$\dot{r} = \mu r + \alpha r^3, \quad \dot{\theta} = \omega. \quad (2.4.34)$$

Suppose that $\alpha < 0$. Then, for $\mu < 0$ we see that $r(t) \rightarrow 0$, since the right hand side of the first equation of (2.4.34) is strictly negative for all $r > 0$. In contrast, for $\mu > 0$ a “radial equilibrium” $r = \sqrt{-\mu/\alpha} > 0$ appears. This is, in fact, not an equilibrium, since the angular variable θ increases (or decreases) continually due to the second equation of (2.4.34). The new solution is a limit cycle. As μ increases through zero a Hopf bifurcation occurs in which the fixed point at $r = 0$ ($(x_1, x_2) = (0, 0)$) loses stability and a stable limit cycle is born. See Fig. 35.

For more on Hopf bifurcations, see [231, §8.4] or for greater detail [89, Chap. 3], which explains how nonlinear coordinate changes can be made that turn any non-degenerate Hopf bifurcation into a standard *normal form* whose leading terms look like those of (2.4.34) when written in polar coordinates. In particular, there is an explicit formula for the coefficient α in (2.4.33-2.4.34) in terms of the first three derivatives of the functions defining the vectorfield $\mathbf{f}(\mathbf{x})$. Normal forms are essentially nonlinear generalizations of the similarity transformations which diagonalize matrices in linear algebra.

Exercise 6. What happens in Eqn. (2.4.33) above as μ passes through zero in the case that $\alpha > 0$? Produce a bifurcation diagram and phase portraits analogous to those of Fig. 35. Explain how the sign of α determines the stability of the limit cycles. Now suppose that a fifth order term is present, so that the radial equation becomes $\dot{r} = \mu r + \alpha r^3 - r^5$, and produce bifurcation diagrams and phase portraits for this case, showing all topologically distinct cases that can occur.

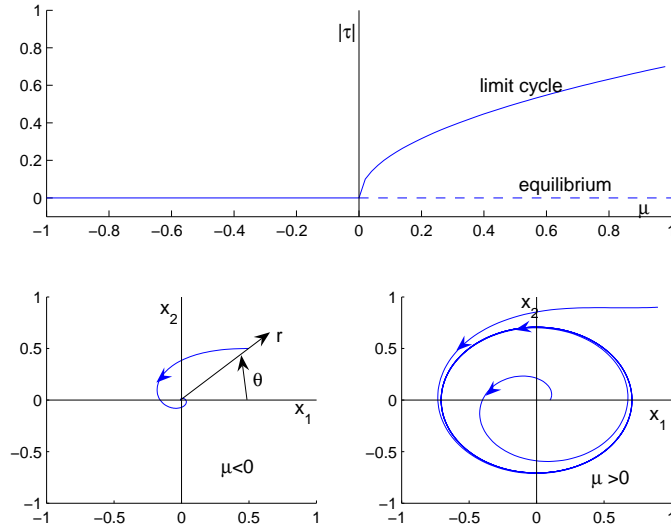


Figure 35: A Hopf bifurcation, showing bifurcation diagram (top) and phase portraits of (2.4.33) for $\alpha < 0$ with $\mu < 0$ and $\mu > 0$ (bottom left and right). The bifurcation diagram shows branches of stable solutions as solid curves and unstable solutions as dashed curves. Note the limit cycle, an attracting periodic orbit, for $\mu > 0$.

2.4.4 Periodic orbits and Poincaré maps

The Hopf bifurcation described above provides a way to deduce the existence of limit cycles or periodic orbits from local information: linearization and computation of a stability coefficient at a degenerate fixed point. Fixed points can occur for ODEs with phase spaces of any dimension $n \geq 1$, but in Euclidean spaces \mathbb{R}^n periodic orbits only appear for dimensions $n \geq 2$. They are *global* phenomena: a solution lying in or attracted to a periodic orbit must leave any sufficiently small neighborhood and circulate around before returning. This clearly cannot occur on the real line. Global behavior is generally hard to detect and analyze for $n \geq 3$ (for example, chaotic behavior can occur [100, 89]), but for planar ($n = 2$) systems there are direct global methods.

The Poincaré-Bendixson theorem [231, §8.1] allows one to prove existence of a periodic orbit “directly” by showing that orbits must enter a certain region that contains no fixed points:

Theorem 3. *Suppose that a planar ODE has an attracting set \mathcal{A} that contains no fixed points. Then \mathcal{A} has a periodic orbit.*

Actually Wilson’s statement is a little more specific and more restricted: it assumes the existence of a positively invariant (attracting) annulus, which implies the presence of at least one asymptotically stable limit cycle. For a more general statement than this or that given above, and a proof, see [100, §10.5]. One can also deduce necessary conditions for the existence of limit cycles using index theory, which considers how tangent vectors rotate as one follows a closed curve [89, §1.8]. This leads to:

Proposition 3. *If a planar ODE has a limit cycle Γ then at least one fixed point must lie inside Γ .*

If multiple points lie within Γ , then the numbers N_s of sources and sinks, and N_c of saddle points, must satisfy $N_s - N_c = 1$.

Hence if one finds a trapping region $\mathcal{B} \subset \mathbb{R}^2$ and hopes to show that a limit cycle lies within it, \mathcal{B} must exclude all fixed points, so it cannot be simply-connected. For example, if there is a single fixed point, \mathcal{B} can be an annulus. Fig. 36 shows an example for the classical van der Pol equation:

$$\begin{aligned}\dot{x}_1 &= x_2 - \left(\frac{x_1^3}{3} - x_1\right), \\ \dot{x}_2 &= -x_1 .\end{aligned}\tag{2.4.35}$$

Finding an inner boundary through which orbits enter the annulus is easy: one need only take a sufficiently small circle around the unstable source at $(x_1, x_2) = (0, 0)$, but it is tricky to construct an outer boundary.

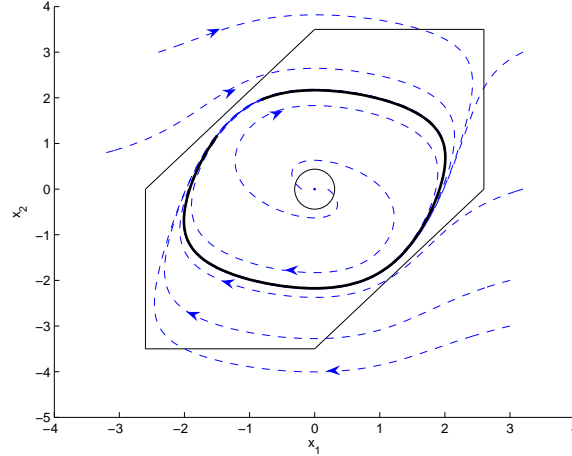


Figure 36: An annular trapping region \mathcal{B} and the limit cycle of the van der Pol equation (2.4.35).

The limit cycle of Fig. 36 is another example of an asymptotically stable or attracting periodic orbit. Periodic orbits can occur in differential equations of any dimension $n \geq 2$, but finding them analytically is much more difficult than finding fixed points, since one must effectively solve the ODE to do it. It is still useful, however, to develop some analytical tools. The *Poincaré map* is one such tool which reduces the analysis of solutions near a periodic orbit γ to the study of a mapping of dimension $n - 1$, one lower than that of the original state space. One defines a (small) piece Σ of an $(n - 1)$ -dimensional manifold (a hyper-plane often suffices), pierced by γ at a point \mathbf{p} and *transverse to the flow* in that the component of the vector field \mathbf{f} normal to Σ does not vanish at any point in Σ . The (hyper-) surface Σ is called a *cross-section*: solutions cross it, all in the same direction, with non-zero speed. By continuity, solutions starting at points $\mathbf{q} \in \Sigma$ near \mathbf{p} follow γ sufficiently closely to intersect Σ again at, say \mathbf{q}' , thus implicitly defining a map on Σ :

$$\mathbf{P} : \Sigma \rightarrow \Sigma \quad \text{or} \quad \mathbf{q} \rightarrow \mathbf{q}' = \mathbf{P}(\mathbf{q}) .\tag{2.4.36}$$

\mathbf{P} is called the *Poincaré* or *first return* map. Note that the point \mathbf{p} at which the periodic orbit intersects the cross-section Σ is a fixed point: $\mathbf{p} = \mathbf{P}(\mathbf{p})$. Poincaré maps are continuous and at least as smooth as the right hand sides of the ODEs that define them.

In the example (2.4.35) above, the positive x_1 -axis

$$\Sigma = \{(x_1, x_2) | x_1 > 0, x_2 = 0\}$$

is a suitable cross-section, and while we cannot integrate (2.4.35) explicitly to obtain a formula for the one-dimensional map P , the instability of the fixed point $(0,0)$ and attractivity of the trapping region \mathcal{B} from outside implies that P takes the form sketched in Fig. 37. It is clear that the continuous function P must intersect the diagonal in at least one point $p > 0$, so that $p = P(p)$ is a fixed point corresponding to the limit cycle. In fact it is true, although not easy to prove, that p is unique and that the linearised map satisfies $0 < (dP/dx_1)|_{x_1=p} = \lambda_p < 1$, implying asymptotic stability of both the fixed point and the corresponding periodic orbit.

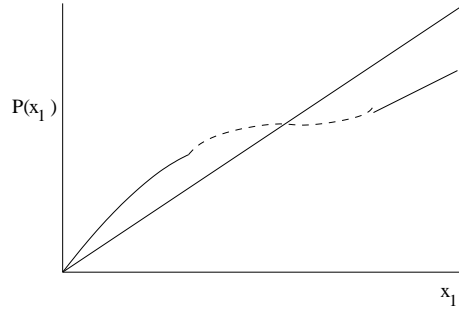


Figure 37: The Poincaré map for Equation (2.4.35).

More generally, from the theory of iterated matrices, if an $(n-1)$ -dimensional Poincaré map $\mathbf{DP}(\mathbf{p})$ linearised at a fixed point \mathbf{p} has only eigenvalues of modulus strictly less than one, then \mathbf{p} and the associated periodic orbit are asymptotically stable. If at least one eigenvalue has modulus greater than one, \mathbf{p} and the periodic orbit are unstable. This parallels the all-negative and at-least-one-positive eigenvalue criteria for flows, described in §2.1. The degenerate case, with one or more eigenvalues of unit modulus, is dealt with rather like the analogous non-hyperbolic fixed point.

As for flows generated by solutions of differential equations, along with the discrete orbits of iterated mappings come invariant manifolds, tangent to the appropriate stable and unstable subspaces of the linearised maps. Suppose that $\mathbf{DP}(\mathbf{p})$ has $s \leq n-1$ eigenvalues of modulus less than one and $u = n-1-s$ of modulus greater than one (for simplicity we assume \mathbf{p} is hyperbolic). Then $\mathbf{DP}(\mathbf{p})$ has, on Σ , s - and u -dimensional stable and unstable subspaces $E^s(\mathbf{p})$, $E^u(\mathbf{p})$ respectively. The stable manifold theorem for maps [89] then yields local manifolds $W_{loc}^s(\mathbf{p})$, $W_{loc}^u(\mathbf{p})$, tangent to $E^s(\mathbf{p})$, $E^u(\mathbf{p})$ at \mathbf{p} , just as for flows. A solution to the ODE started on Σ in $W_{loc}^s(\mathbf{p})$ will circulate near γ and return to a point of $W_{loc}^s(\mathbf{p})$ closer to \mathbf{p} at its next intersection with Σ . All such solutions started in a neighborhood $B(\mathbf{p}) \subset \Sigma$ form the $(s+1)$ -dimensional local stable manifold of the periodic orbit γ , $W_{loc}^s(\gamma)$, while solutions started at points of $W_{loc}^u(\mathbf{p})$ form the $(u+1)$ -dimensional local unstable manifold of γ , $W_{loc}^u(\gamma)$. Fig. 38 shows two possible structures in the three-dimensional case in which $s = u = 1$. In case (a) the real eigenvalues of $\mathbf{DP}(\mathbf{p})$ satisfy $0 < \lambda_1 < 1 < \lambda_2$; in case (b) $\lambda_2 < -1 < \lambda_1 < 0$. In the latter case it is easy to see that the (diagonalised) matrix

$$\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

of $\mathbf{DP}(\mathbf{p})$, which takes $(x_1, x_2) \in \Sigma$ to $(\lambda_1 x_1, \lambda_2 x_2)$, rotates the vector connecting $(0,0)$ to (x_1, x_2) by (approximately) π . As solutions circulate in a tubular neighborhood of γ , therefore, they pick

up an odd number of half twists. In case (a) the number of half twists is even. Fig. 38 depicts the simplest cases of 0 and 1 half twists respectively. In the latter case the local stable and unstable manifolds are Möbius bands. Starting on the right of the periodic orbit, say, and following it on either manifold for one circuit, one returns on its left.

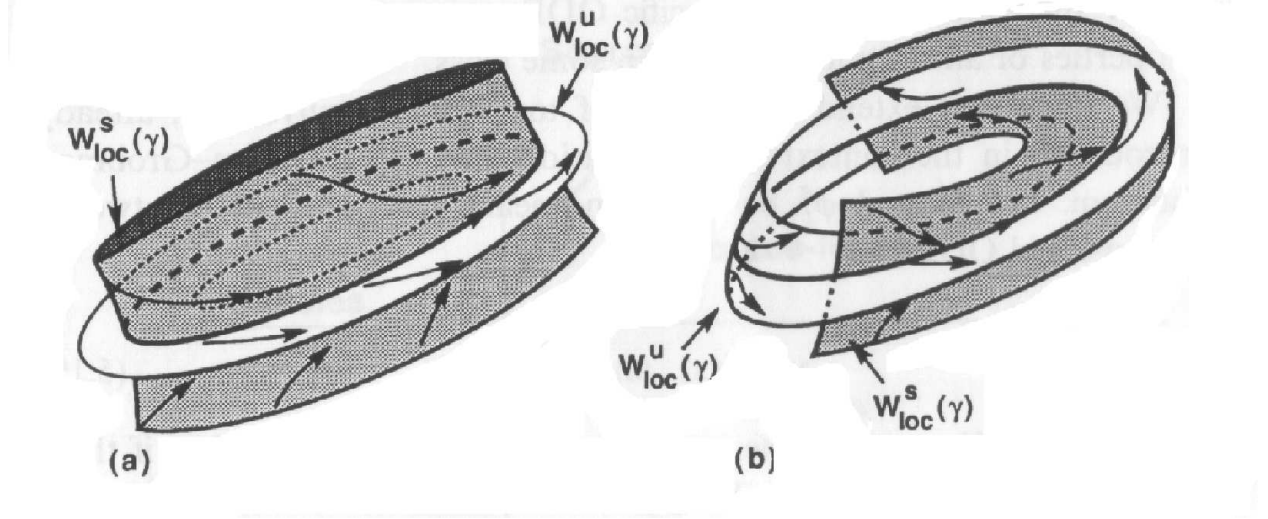


Figure 38: Periodic orbits in a three-dimensional flow with (a) orientable and (b) non-orientable stable and unstable manifolds. The stable manifolds are lightly shaded and in (b) part of $W_{loc}^s(\gamma)$ is removed for clarity. Figure taken from [108, Fig. 5.4].

2.5 An introduction to neural networks as nonlinear ODEs

In this section we present two abstracted examples of neural networks and their realizations as systems of ODEs. These *connectionist* or *firing rate* models [185, 232, 233, 97] ignore the details of spike dynamics and synaptic transmission, replacing the fine scale temporal dynamics of transmembrane voltages by variables describing the activations or firing rates of cells or groups of cells. They are considerably simpler and more tractable than the Hodgkin-Huxley equations (which we introduce in §§3.3-3.4).

Inhibitory connections lower a postsynaptic neuron's activation in response to presynaptic activity; this is accomplished in ODEs by negative coefficients multiplying the presynaptic activation terms in the postsynaptic equation. Similarly, positive coefficients represent excitatory connections. Stimuli originating in sensory neurons can thereby either increase or decrease firing rates depending on the neuronal response properties, but the effects are typically nonlinear.

In firing rate models, postsynaptic neural responses to inputs from other cells, and to stimuli, are characterized by simple input/output functions or current-to-frequency transduction curves. Firing rates are bounded: they cannot drop below zero, and they cannot exceed a maximum frequency, controlled by the refractory period, and depending on the cell type. Thus, low level stimuli typically have little effect, while the response saturates at high stimulus levels. Between these limits the cell is maximally responsive to changes in stimuli. This intuitively justifies the use of sigmoid-like input/output functions. Such functions can be derived from spiking neuron models by mean field

and kinetic theory methods [176, 198], as described in §6.2 below.

The size and scales of the “compressive effect” vary from system to system, so several functional descriptions have been developed. In addition to the sigmoid of Eqn. (2.4.28), the *Naka-Rushton function* describes the response $S(P)$ to a stimulus of strength P via three parameters:

$$S(P) = \frac{MP^N}{\sigma^N + P^N} \quad \text{for } P \geq 0 \quad \text{and} \quad S(P) = 0 \quad \text{for } P < 0. \quad (2.5.1)$$

Here $S(P) \rightarrow M$, the maximal response (spike rate) for strong stimuli. The parameter σ sets the stimulus strength at which $S(P)$ reaches $M/2$ and N controls the steepness with which the rise from 0 to M occurs. In the sigmoid of (2.4.28) the parameters g (gain) and β (bias, or offset) play similar roles to N and σ in (2.5.1) (the maximum value was normalized to 1 in (2.4.28)). Piecewise linear functions are also sometimes used [214, 25].

The following ODEs describe a simple two neuron network with the winner-take-all property [231, §6.7]:

$$\frac{dE_1}{dt} = \frac{1}{\tau}(-E_1 + S(K_1 - 3E_2)), \quad (2.5.2a)$$

$$\frac{dE_2}{dt} = \frac{1}{\tau}(-E_2 + S(K_2 - 3E_1)) : \quad (2.5.2b)$$

see Fig. 39. Here E_1 represents the spike rate of neuron 1, which receives external input K_1 , and is inhibited by neuron 2, and E_2 represents the spike rate of neuron 2, which neuron 1 inhibits. Although we say “neuron 1 and neuron 2,” each of these mathematical variables may represent the average activity over a *population* of cells of the same type, or a subgroup tuned to a particular stimulus (as in the Usher-McClelland model of Example 7).

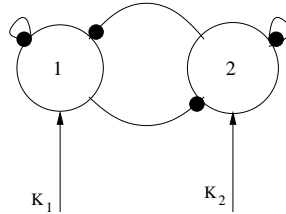


Figure 39: A simple network with winner-take-all dynamics. Neurons or groups of neurons are represented as circles, and external inputs and synapses to other neurons as connecting lines or arcs. Excitatory connections (positive inputs) terminate with an arrow and inhibitory connections (negative inputs) with a filled circle. Self-inhibitory connections are sometimes omitted in such depictions. The self-excitatory connections are absent in Eqns. (2.5.2). **FIG NEEDS FIXING: CHANGE dots on self connections to arrows!**

For suitable parameter values (2.5.2) has two sinks whose domains of attraction are separated by the stable manifold of a saddle point. In particular, for $M = 100$, $\sigma = 120$, $N = 2$ and $K_1 = K_2 = 120$, they lie at $(0, 50)$ and $(50, 0)$. This is the essence of the winner-take-all idea. If neuron 1, say, has a higher firing rate, either from the initial conditions or due to a stronger input, its output drives down that of neuron 2, thus decreasing inhibitory feedback to 1 and allowing 1's rate to increase further. This continues until neuron 2's state is driven to zero. This system, which is explored further in Exercise 7 below, is a simple model for deciding between two alternatives

based on evidence (external inputs). If the solution goes to the equilibrium with neuron 1 non-zero, then the system “chooses” alternative 1. If it goes to the equilibrium with neuron 1 driven to zero, then it chooses 2. Note that the decision is determined both by the inputs (evidence for each choice) *and* by initial conditions (e.g., bias towards one alternative). Biases that set the system close to an equilibrium can require strong evidence to converge on the correct decision! Also, if inputs are changed after the system makes a decision, it is likely to make the same choice in the next trial. We take up decision models in §6.

Exercise 7. *Plot nullclines for the ODEs (2.5.2), find all the equilibria and determine their stability types for the parameter values specified above (you may additionally take $\tau = 20$: how does this time constant affect stability?). Repeat for $K_1 = 130$, $K_2 = 110$ and for $K_1 = 150$, $K_2 = 90$ with the other parameters unchanged. In each case plot nullclines and solution trajectories for five different initial conditions. In the decision-making context, what is significant about the third case?*

We now turn to a multi-stage, higher-dimensional system: a neural network describing retinal light adaptation. See [231, §7.3] for further details and an illustrative experiment that allows you to experience your own time constants. The network is illustrated in Fig. 40, in which the letters identify types of neurons, and, in the ODEs (2.5.3) below, represent the spike rates of those cells (or populations of them).

Light reaching the retina activates cone photoreceptors (C), which excite bipolar cells (B) and horizontal cells (H). The bipolar cells excite amacrine cells (A) and ganglion cells (G). Ganglion cell axons conduct signals further into the visual system, which is not modeled here. Horizontal cells inhibit the cones with subtractive feedback and the bipolar cells with feedforward inhibition. The amacrine cells excite the interplexiform cells (P), which excite the horizontal cells. Finally, amacrine cells inhibit the bipolar cells with divisive feedback, as in Example 6 above (albeit with a different constant premultiplying A in the denominator). As there, in the winner-take-all network above, and in Example 7, spike rates are assumed to decay exponentially in the absence of inputs. This model incorporates the “H - C” model of [231, §3.3] and the “A - B” model of Example 6 and unites them in a more general setting.

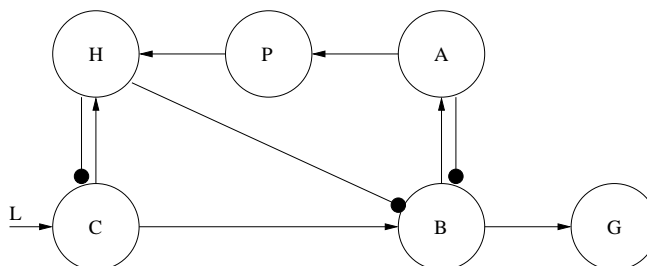


Figure 40: The retinal light adaptation circuit modeled in Eqns 2.5.3.

This six-unit network is modeled by the following ODEs, where L again represents intensity of

light incident on the retina:

$$\frac{dC}{dt} = \frac{1}{10}(-C - PH + L), \quad (2.5.3a)$$

$$\frac{dH}{dt} = \frac{1}{100}(-H + C), \quad (2.5.3b)$$

$$\frac{dB}{dt} = \frac{1}{10} \left(-B + \frac{6C - 5H}{1 + 9A} \right), \quad (2.5.3c)$$

$$\frac{dA}{dt} = \frac{1}{80}(-A + B), \quad (2.5.3d)$$

$$\frac{dP}{dt} = \frac{1}{4000}(-P + 0.1A), \quad (2.5.3e)$$

$$\frac{dG}{dt} = \frac{1}{10} \left(-G + \frac{50B}{13 + B} \right). \quad (2.5.3f)$$

The values for the time constants and other parameters are taken from [231]. The Naka-Rushton function appears again in the equation for G , with $N = 1$; this function was originally invented to describe retinal ganglion cells.

We start our analysis by finding equilibria and analyzing their stability. Equilibria satisfy

$$C = -PH + L, \quad (2.5.4a)$$

$$H = C, \quad (2.5.4b)$$

$$B = \frac{6C - 5H}{1 + 9A}, \quad (2.5.4c)$$

$$A = B, \quad (2.5.4d)$$

$$P = 0.1A, \quad (2.5.4e)$$

$$G = \frac{50B}{13 + B}. \quad (2.5.4f)$$

We may immediately eliminate the three variables H , A and P in favor of B and C , using $H = C$, $A = B$, and $P = 0.1A$. Moreover, none of the other variables depends on G (Eqn. (2.5.3f) decouples), and G depends only on B , so we can solve for C as a function of B and then form a single equation which can be solved for B :

$$C = \frac{10L}{10 + B}, \quad B = \frac{C}{1 + 9B} \Rightarrow 9B^3 + 91B^2 + 10B = 10L. \quad (2.5.5)$$

Given a light intensity $L \geq 0$, this cubic polynomial has a unique positive root in $B \geq 0$ (why?), and solving it (numerically?) we may substitute back into (2.5.4) to get the equilibrium values for the other five variables. From (2.5.5) we note that $B \approx L$ for small L , while for large L , $B \approx L^{1/3}$: the cube root nonlinearity effectively compresses the ganglion cell's output range at high intensities.

Exercise 8. *Derive the Jacobian matrix for this equilibrium in terms of L , and show that it is stable for $L = .01, .1, 1, 10, 100, 1000$. [Hint: use `eig()` in MATLAB].*

The intermediate stages are critical to light adaptation. If L were input directly to the ganglion cells G in place of B , then a change from intensity L to $L + \Delta$ would produce an incremental change in ganglion firing rate of

$$\delta G = \frac{50(L + \Delta)}{13 + (L + \Delta)} - \frac{50L}{13 + L}. \quad (2.5.6)$$

Supposing that $\delta G = 1$ (in arbitrary units) is the smallest discriminable change that can be recognized, and assuming that $\Delta \ll L$, we have

$$\begin{aligned}
\delta G = 1 &= \frac{50(L + \Delta)}{13 + (L + \Delta)} - \frac{50L}{13 + L} \\
\Rightarrow (13 + L + \Delta)(13 + L) &= 50[(13 + L)(L + \Delta) - (13 + L + \Delta)L] \\
\Rightarrow (13 + L)(L + \Delta - L) - \Delta L &= \frac{(13 + L + \Delta)(13 + L)}{50} \\
\Rightarrow \Delta &= \frac{(13 + L + \Delta)(13 + L)}{50 \times 13} \approx \frac{(13 + L)^2}{50 \times 13}.
\end{aligned} \tag{2.5.7}$$

Hence, for large L , the smallest discriminable intensity change increases like L^2 , making intensity differences of bright lights hard to distinguish. Sensitivity is drastically improved by adaptation, which relies on the relatively slow dynamics occurring in the intermediate processing stages of Fig. 40.

Equation (2.5.3) has three distinct time scales: C , B , and G have time constants of 10 ms, which is much faster than those of the other variables, A and H have time constants of 80 and 100 ms, respectively, and P has a very slow time constant of 4000 ms. Hence, following a sudden change in L , the variables C , B , and G rapidly approach what would be a stable equilibrium if H , A and P were fixed. But they are not, and A and H change on the intermediate timescale, while P begins its slow change. This leads to the behavior illustrated in Fig. 41, where the three timescales are apparent: a rapid rise in B and G , a slightly slower rise in A , followed by equilibration of these three variables to the new equilibrium, while P continues to approach $0.1A$ very slowly.

Exercise 9. Write a MATLAB program simulating the system in Equation 2.5.3 for a given level of light L . Incorporate a way to have the light step to a higher intensity during the simulation. Explore for different starting intensities L what increment Δ is required so that the spike in G is of magnitude 1. Allow the system to reach steady state before you increase the intensity. Do NOT record the new steady state adapted value of G , but measure the height of the transient spike. Find Δ for L initially equal to 1, 2, 10, 20, 100, 200, 1000, 2000, 10000, and plot Δ versus L on a logarithmic scale. Also plot the value of Δ from Equation (2.5.7), which does not incorporate the adaptation mechanism (cf. [231, Fig. 7.8]).

Although the timescale separation is not really sufficient, we can give an idea of how it may be used to reduce the dimension of the system, a method that will be important in §3. Since C , B and G equilibrate rapidly, we *assume* that they instantaneously attain the values specified in Eqns. (2.5.3a), (2.5.3c) and (2.5.3f) as they track the other variables. Substituting into the remaining ODEs for H , A and P , we obtain a three-dimensional system that describes behavior on the 100 to 1000 ms time scale. Furthermore, since P changes very slowly, the essential dynamics on the intermediate 100 ms time scale is governed by the two-dimensional system

$$\frac{dH}{dt} = \frac{1}{100}(-H - PH + L) = \frac{1}{100}(-(1 + P)H + L), \tag{2.5.8a}$$

$$\frac{dA}{dt} = \frac{1}{80} \left(-A + \frac{6(L - PH) - 5H}{1 + 9A} \right) = \frac{1}{80} \left(-A + \frac{6L - (5 + 6P)H}{1 + 9A} \right). \tag{2.5.8b}$$

This system changes as P , acting like a slowly-varying parameter, moves to its new equilibrium. Here no bifurcations occur, but they can do so in other systems, and we shall use reduction methods of this type along with bifurcation theory to elucidate the dynamics of bursting in §3.5. In general,

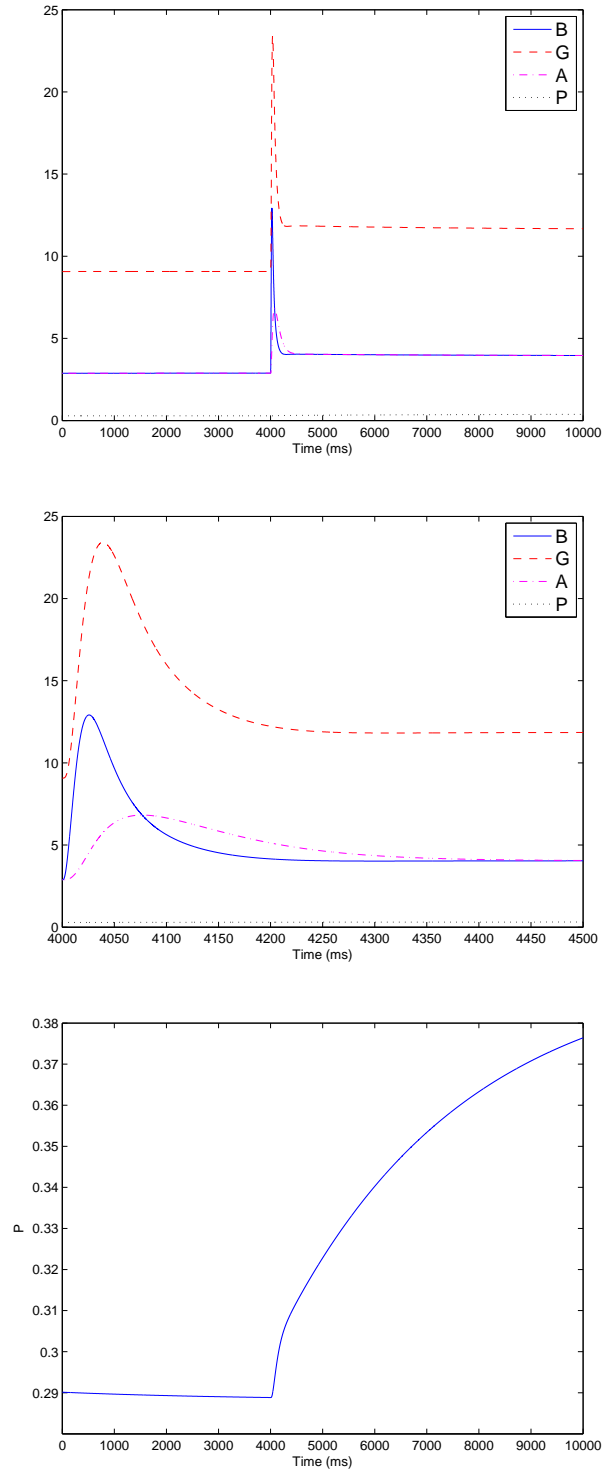


Figure 41: Solutions of Eqns 2.5.3 modeling retinal adaptation following an intensity step from $L = 100$ to $L = 200$. Top and middle: A , B , P and G over 10 sec, and over 500 ms immediately following the stimulus change. Bottom: P over 10 sec with expanded vertical scale.

identifying the relevant time scales in a system, modeling at the appropriate time scales for the phenomena of interest, and justifying the reductive methods, are all major research challenges.

Exercise 10. *Find the equilibrium for Eqn. (2.5.8) as a function of L and P , and show that it is stable.*

Summarizing, cells C , B , and G form a direct excitatory pathway from L , but inhibition of B from H and A , modulated by P , introduces the adaptive dynamics. Mechanisms such as this allow sensory systems to be sensitive to small changes over broad ranges of stimuli. For an illustrative experiment using your own retinas, see [231, Fig. 7.9, p 103].

3 Models of single cells (≈ 2 weeks)

Having developed some of our mathematical toolbox, we now return to the main topic of the course: neurobiological modeling. Unlike the classic early books of Wiener [229] and von Neumann [217] on modeling the nervous system and brain, and the models introduced in §2.5 above, in this and the following section we shall take a reductive, cellular-level viewpoint. We start with general observations about models.

3.1 Modeling strategies and scales

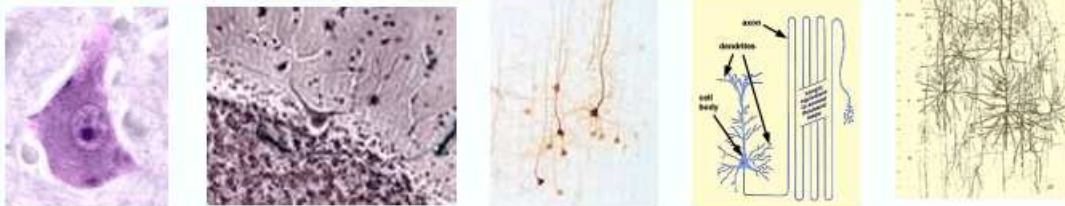
There are many approaches to the mathematical modeling of biological phenomena, each with different advantages and uses. This course explores several of these avenues. The first and most obvious characteristic of a model is the level or *scale* at which it operates (cf. [231, §1.2]). In ecology the focus is often on populations and little detail is accorded to individual organisms beyond attributes such as infected or susceptible, male or female, age or developmental stage, etc. While there is considerable interest in communities of individuals engaged in collective decision-making (in neuro-economics, for example), neurobiological models usually start at the level of individuals, with full nervous system models, which identify the major nerves connecting parts of the body and various organs to the central nervous system, as described in §1.3.

The next level of detail focuses on the brain, modeling interactions among regions or brain areas within it. Zooming in further, one can subdivide each area into smaller groups of neurons, for example: those tuned to respond to particular elements of stimuli such as horizontal bars, and subgroups of those tuned for vertical bars (Fig. 42). At these levels there is no concern for the membrane potential of a single neuron or its morphology; the scale is still too large to be concerned with details that do not (obviously) influence the behavior. Continuing to smaller scales, we progress to models of several neurons, at which point spike rates, membrane voltages, and phases of individual cells are of interest. But even these models do not address the anatomical or physiological detail of single cells.

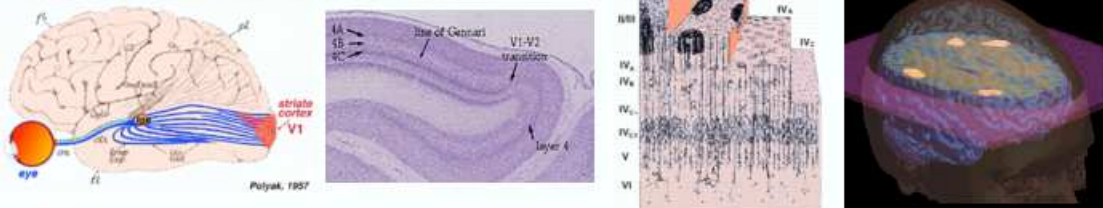
At the next level of detail, where the Hodgkin-Huxley model resides [231, Chap. 9], the neuron is viewed as a single compartment containing different ions that drive transmembrane currents. Here

The multiscale brain:

Ingredients: $\sim 10^{11}$ neurons, $\sim 10^{14}$ synapses.



Structure: layers and folds.



Communication: via action potentials, spikes, bursts.

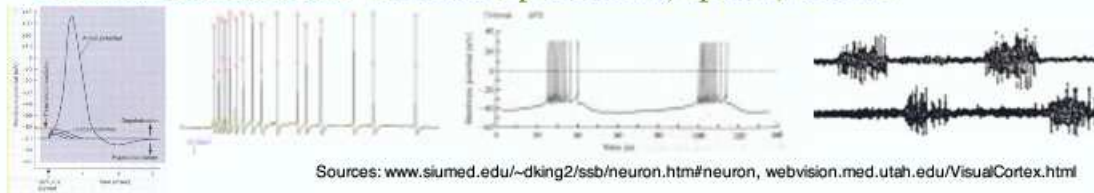


Figure 42: Phenomena in the brain exist at many different spatial and temporal scales.

ionic currents and membrane potentials characterize the model, but although it is far more detailed than those at larger scale, it is still imprecise and crude in many ways. Variations in properties along axons and dendritic processes of cells that make them functions of *space* as well as time are absent, and the morphology of the neuron is ignored. At the next level details of dendritic trees, axons and synapse locations can be added, but even these ignore the biochemical and molecular machinery that governs each ion channel and determines neurotransmitter release and transport across the synaptic cleft.

Thus far we have been discussing spatial scale, from body size to molecules, but temporal scales play equally important roles. Spatial and temporal scales usually go hand in hand, and as illustrated in Figure 43 and explored in more depth later, the range of temporal scales in the brain actually facilitates mathematical modeling and analysis, since it allows one to separate phenomena that take place at different rates and treat them quasi-independently, approximating fast ones as instantaneously equilibrated and slow ones as fixed. The Hodgkin-Huxley equations, treated in this chapter, describe phenomena that last for milliseconds, and synapses work at a similar scale. The time scale increases to hundreds of milliseconds to seconds when considering behavioral responses to stimuli. Decision making takes place over fractions of a second to minutes. Learning includes short- and long-term effects, with time scales ranging from minutes through hours and days to years. Developmental neuroscience studies the formation of the embryonic nervous system, and its relevant time scale can be months to years, depending on the species.

Multiple scales in the brain and in math:

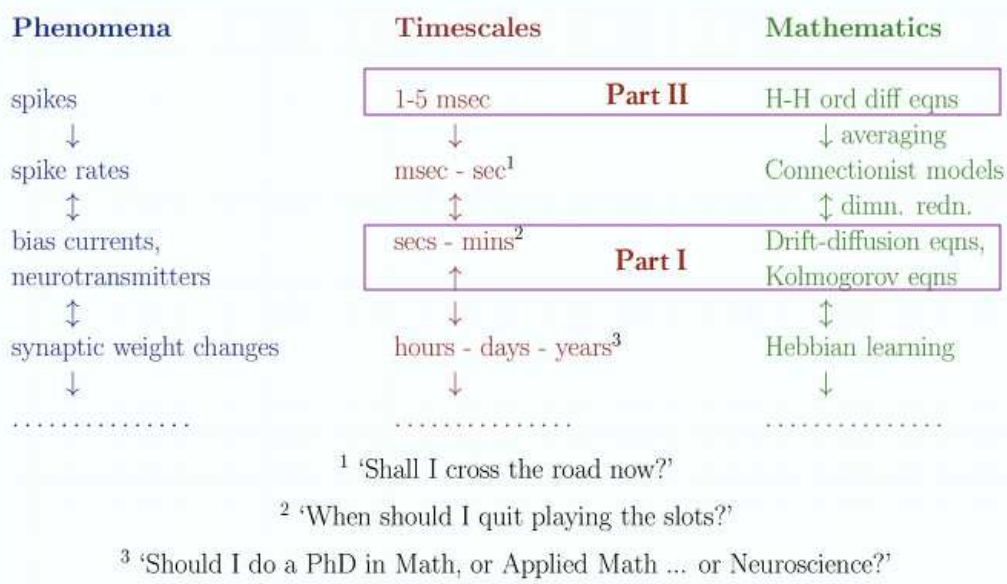


Figure 43: The range of time scales is addressed by a range of mathematical models (taken from an introductory lecture of PH: in this course Part II refers to §§3-4 and part I to §6).

At any given level of detail, there are several types of models that can be developed. As noted in §1.1, Mechanistic models aim to accurately capture the underlying physiology, and thereby make predictions about physiological behavior which can be detected experimentally. Models of this type, while seemingly the best, are often 1) complicated (nonlinear and high-dimensional), thereby limiting useful analysis, 2) extensive, limiting large scale simulation of linked models, 3) characterized by many parameters which are difficult to measure or estimate, and 4) therefore rare in practice. Other models simplify the physiology, often only focussing on one feature; they consequently have reduced dimensionality compared to more complete models, and are therefore often less useful for prediction and explanation. Finally, at the far end of the spectrum, are empirical or phenomenological models that attempt to fit experimental data, without reference to underlying physiology. Typical models rarely fall cleanly into one of these three classes; they inhabit a continuum, with most somewhere near the middle. As these notes proceed, we will develop and analyse a representative set of them.

Fortunately, for our first major exercise, the Hodgkin-Huxley (H-H) model is excellent in several aspects. It is a cellular-level model which was established by careful physiological experiments (although it is crude when viewed from the molecular scale). It and its modifications and extensions have successfully described a wide variety of observations. For instance, years after their development in the early 1950's, it was noticed that the H-H equations predicted a phenomenon known as hysteresis, which had never been seen experimentally. In 1980 hysteresis was verified in experiments of Guttman, Lewis and Rinzel [91], cf. [231, p. 142]. This example demonstrates the power of a physiologically-inspired model.

Other models that we will see in this section are simplifications of H-H, which retain physiological inspiration but are less accurate, less detailed and more amenable to analysis. The first is a two-dimensional reduction due to Rinzel [180]. The second, the FitzHugh-Nagumo (F-N) model [64, 158], predates Rinzel's and was not directly inspired by physiology, but by examining H-H solutions and creating a more tractable two dimensional ODE with similar dynamics. While not directly modeling physiology, it is still very useful, since it identifies key *qualitative* properties that produce the observed action potential (AP) or spike and the refractory behavior that follows it: properties that other more detailed models share.

If one is interested in, e.g., the accumulation of spiking activity in an oculo-motor area preceding a saccade in response to a visual stimulus (see §6), one would prefer to model a single variable – the short-term average spike rate in that region – rather than the details of membrane potentials and ionic transport in every neuron that determine it. Thus, a major challenge is extracting a closed model at the appropriate scale to describe the phenomenon of interest. Hodgkin and Huxley achieved this at the cellular scale by modeling ion channels phenomenologically, fitting functions that describe the voltage-dependence without reference to the (then unknown) molecular detail involved in gating the channels.

In the remainder of this section we describe the H-H model, and then consider simpler models for which mathematical analyses can go further. We go beyond APs in single cells to examine propagation of APs, bursting effects and communication between neurons at synapses. In the following section (§4) we describe synaptic connections and interactions among multiple neurons in small neural circuits that produce repetitive rhythms: central pattern generators (CPGs). However, we must begin at a more fundamental level and smaller scale than a single neuron, which turns out

not to be so simple.

More extensive accounts of parts of the material in this and the next section can be found in [120, 124, 61].

3.2 Ion channels and membranes

All animal cells are enclosed by a membrane consisting primarily of a lipid bilayer about 7.5 nm thick. Its primary purpose is to separate the inside of the cell from what is outside, but it also allows passage to various molecules, through many pores and channels. The membrane itself apart from the pores and channels prevents the unregulated passage of water, and of sodium, potassium, and chloride ions, and therefore resists the easy flow of electrical current. The membrane exhibits electrical resistance and capacitance, each of which are modified by the states of the channels. Different channels allow the movement of different ions; hence we refer to sodium channels, potassium channels, etc.

Ionic transport is important for many cell processes, but is *essential* for neural phenomena. Ions and molecules can be moved by both active and passive processes. Water crosses the membrane passively through osmosis, which is controlled via ion concentrations, allowing the cell to regulate its volume. Ions also cross the membrane passively via diffusion through pores. Lipid-soluble molecules such as carbon dioxide diffuse through the lipid bilayer itself. Differences in ion concentration drive water osmotically, and create an electrical potential across the membrane, known as the *membrane potential*. Finally, ionic concentration differences tend to decrease due to diffusion, creating the need for active processes to balance the passive ones.

Active processes include pumps that exchange sodium in the cell for potassium and pumps that remove calcium from the cell. For example, after cell death when active process of calcium Ca^{2+} removal stops, the Ca^{2+} concentration in the cell rises much higher than in a living cell, keeping the cell in the constant tension of rigor mortis [124, Chap. 5]. One pump exchanges three sodium Na^+ ions for two potassium K^+ ions, maintaining the intracellular K^+ concentration much higher than in the extracellular space, and the extracellular Na^+ concentration higher than in the cell. The pump works against concentration gradients, and so requires energy (in the form of ATP) to operate.

Since diffusion plays such an important role in determining intracellular ionic concentrations, we briefly examine a mathematical model of it [231, Chap. 15]. This discussion will also introduce another kind of mathematical model: a *partial differential equation* (PDE). Given a region of space Ω , we denote by $c = c(\mathbf{x}, t)$ the concentration of the ion of interest as a function of space and time over Ω . Letting q be the production of c per unit volume defined over Ω and \mathbf{J} the vector flux of c defined along the boundary of Ω , with \mathbf{n} the unit normal to the boundary, we obtain the following conservation law:

$$\frac{\partial}{\partial t} \int_{\Omega} c dV = \int_{\Omega} q dV - \int_{\partial\Omega} \mathbf{J} \cdot \mathbf{n} dA. \quad (3.2.1)$$

(In words: the rate of change of c in Ω = production – loss through boundary: this is an example

of a *conservation law*.) By the divergence theorem, we have

$$\int_{\partial\Omega} \mathbf{J} \cdot \mathbf{n} dA = \int_{\Omega} \nabla \cdot \mathbf{J} dV, \quad (3.2.2)$$

so that for a fixed region Ω (3.2.1) becomes

$$\int_{\Omega} \frac{\partial c}{\partial t} dV = \int_{\Omega} (q - \nabla \cdot \mathbf{J}) dV. \quad (3.2.3)$$

This integral conservation law holds for any fixed region Ω , and since the region is arbitrary, the integrand must be identically zero. We therefore obtain a partial differential equation (PDE) describing the rate of change of c :

$$\frac{\partial c}{\partial t} = q - \nabla \cdot \mathbf{J}. \quad (3.2.4)$$

To close this equation, we need an expression for the ion flux \mathbf{J} , in terms of c . Fick's law, which is not a natural law, but a good approximation under many circumstances (similar to Ohm's Law), states that flux is proportional to the negative of the concentration gradient. This matches our intuitive understanding that ions tend to flow from regions of high concentration to those of low concentration. Substituting Fick's law with diffusion coefficient D into (3.2.4), we obtain

$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) + q. \quad (3.2.5)$$

Specializing to a single space dimension ($\mathbf{x} = x$), (3.2.5) becomes

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial c}{\partial x} \right) + q, \quad (3.2.6)$$

and if D is constant, the right hand side may be written $D \frac{\partial^2 c}{\partial x^2}$: the classical “heat equation” first proposed by Fourier to describe heat conduction. Note that in steady state, with no local production (eg: release of bound ions such as Ca^{2+} from the endoplasmic reticulum), the concentration profile is linear. This is easily checked.

Random, passive diffusion and active pumps are not the only processes affecting ion concentrations. Positive ions tend to move from high to low electrical potentials. Let z be the valence of the ion, so the quantity $\frac{z}{|z|} = \pm 1$ denotes the sign of the charge on the ion. As above, c is the concentration of the ion, and u describes its mobility: a different constant for each ionic species. Finally, ϕ denotes the electrical potential, and recall from physics that a potential gradient is the same as an electrical field. Planck's equation describes the ion flow resulting from such a potential gradient [124, Chap. 2.6]:

$$\mathbf{J} = -u \frac{z}{|z|} c \nabla \phi. \quad (3.2.7)$$

There is thus a flux due to concentration gradients and a flux due to potential gradients. Each is governed by a constant specific to an ion. Einstein realized that the constants for the same ion were related, and developed a relationship between the diffusion constant D in Fick's law, and the ionic mobility u in Planck's equation [124, p. 53]. If F is Faraday's constant, T is the absolute temperature, and R is the universal gas constant, this reads:

$$D = \frac{uRT}{|z|F}. \quad (3.2.8)$$

Combining the concentration-driven diffusion (3.2.5) with the Planck's equation flux (3.2.7) using Einstein's relationship (3.2.8) yields the Nernst-Planck equation

$$\mathbf{J} = -D \left(\nabla c + \frac{zF}{RT} c \nabla \phi \right). \quad (3.2.9)$$

From this equation, which relates ion flux, concentration difference, and potential difference, we can now derive the *Nernst potential*: the transmembrane potential difference consistent with zero flux for a given concentration difference. We model the membrane as one-dimensional without spatial variations parallel to the boundary: Fig. 44. Partial derivatives in (3.2.9) become one-dimensional spatial derivatives, gradients also become spatial derivatives, as in (3.2.6), and setting the flux $\mathbf{J} = \mathbf{0}$ allows one to solve for the potential difference:

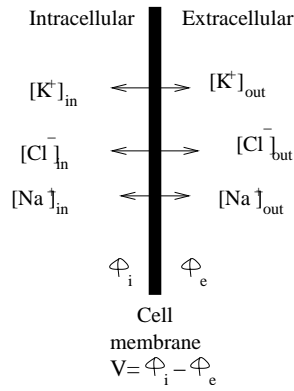


Figure 44: Schematic of a 1-d cell membrane showing ion concentrations and membrane potential.

$$-D \left(\frac{dc}{dx} + \frac{zF}{RT} c \frac{d\phi}{dx} \right) = 0 \quad \Rightarrow \quad \left(\frac{1}{c} \frac{dc}{dx} + \frac{zF}{RT} \frac{d\phi}{dx} \right) = 0. \quad (3.2.10)$$

Integrating (3.2.10) across the membrane length, with c_e and c_i denoting the external and internal concentrations and ϕ_e and ϕ_i the potentials, we obtain

$$\ln(c) \Big|_{c_i}^{c_e} = \frac{zF}{RT} (\phi_i - \phi_e). \quad (3.2.11)$$

This equation relates concentration differences to potential differences for the zero flux condition, which is the condition for the Nernst equation. Recognizing the difference $V = \phi_i - \phi_e$ as the membrane potential and exponentiating (3.2.11) yields the Nernst equation:

$$V = \frac{RT}{zF} \ln \left(\frac{c_e}{c_i} \right) : \text{ the Nernst potential.} \quad (3.2.12)$$

However, even if one ionic species is in equilibrium according to (3.2.12), others will typically not be, thus creating nonzero fluxes and currents that can change the potential. The Nernst equation (3.2.12) is derived from first principles of thermodynamic equilibrium and does not depend on our model, but to obtain the potential for zero net electrical current, we must model the current-voltage relationship for the cell membrane. We have no first-principles argument to derive the current, and indeed this $I - V$ relationship differs for different cells, so here we describe two simple models that cover a wide range of cases.

The *Goldman-Hodgkin-Katz (GHK)* current equation is obtained from the Nernst-Planck equation [124, p. 55] as follows. Assuming a constant electric field across a cell membrane of thickness L and potential V , the field within the membrane is $E = -d\phi/dx = -V/L$. This is then substituted into (3.2.9) to obtain

$$\frac{dc}{dx} - \frac{zFV}{RTL}c + \frac{J}{D} = 0, \quad (3.2.13)$$

which we can solve for $c(x)$. Note that we may take $J(x) = J$ as constant, since the current does not change across the membrane, for there is no local accumulation of charge. The boundary conditions are that $c(0) = c_i$ and $c(L) = c_e$, and using the former we get:

$$\exp\left(\frac{-zVFx}{RTL}\right) c(x) = \frac{JRTL}{DzVF} \left[\exp\left(\frac{-zVFx}{RTL}\right) - 1 \right] + c_i. \quad (3.2.14)$$

So far we have only used the intracellular boundary condition $c(0) = c_i$ but we may now choose the transmembrane current J such that $c(L) = c_e$. We also rename D/L as the parameter P_S : the permeability of the membrane to ion S , and setting $I_S = zFJ$ we replace J by I_S : the GHK current for ion S . Here J is the flux in moles per unit area per unit time and zF is the charge carried per mole of ion S . Thus I_S is the charge per unit area per unit time, or the current per unit area:

$$I_S = P_S \frac{z^2 F^2}{RT} V \frac{c_i - c_e \exp\left(\frac{-zFV}{RT}\right)}{1 - \exp\left(\frac{-zFV}{RT}\right)}. \quad (3.2.15)$$

Note that $J = 0$ and hence the GHK current (3.2.15) is also zero at

$$V = \frac{RT}{zF} \ln\left(\frac{c_e}{c_i}\right), \quad (3.2.16)$$

which is the Nernst potential, so the GHK ionic current model is consistent with our work from first principles.

Exercise 11. Solve Equation 3.2.14 for $c(x)$, with the given boundary conditions, and verify that using Equation (3.2.15) in Equation (3.2.14) results in $c(L) = c_e$.

We now wish to determine the membrane potential at which the *net* ionic current is zero: the *resting potential*. Obviously, for given intra- and extracellular concentrations, each ion has a Nernst potential given by (3.2.12), but how do we get the overall membrane resting potential? The GHK current equation (3.2.15) allows us to add the currents and obtain the requisite balance. With some ions with valence $z = 1$ and others with valence $z = -1$, we get zero total current if

$$0 = \sum_{z=1} P_j \frac{c_i^j - c_e^j \exp\left(\frac{-VF}{RT}\right)}{1 - \exp\left(\frac{-VF}{RT}\right)} + \sum_{z=-1} P_j \frac{c_i^j - c_e^j \exp\left(\frac{VF}{RT}\right)}{1 - \exp\left(\frac{VF}{RT}\right)}, \quad (3.2.17)$$

which allows us to solve for the GHK potential V :

$$V = \frac{RT}{F} \ln\left(\frac{\sum_{z=-1} P_j c_i^j + \sum_{z=1} P_j c_e^j}{\sum_{z=-1} P_j c_e^j + \sum_{z=1} P_j c_i^j}\right). \quad (3.2.18)$$

We can now express the resting potential of a neuron in terms of concentrations and permeabilities. Remember that the GHK derivation requires the constant electric field assumption, whose validity depends on the type of cell, the concentration levels, and other factors.

Exercise 12. Given that $RT/F \approx 25.8$, and that for the squid giant axon the intracellular and extracellular concentrations of the ions are $[K^+]_{in} = 397$, $[K^+]_{out} = 20$, $[Na^+]_{in} = 50$, $[Na^+]_{out} = 437$, $[Cl^-]_{in} = 40$, and $[Cl^-]_{out} = 556$, and that the permeability ratios are $P_K : P_{Na} : P_{Cl} = 1 : 0.03 : 0.1$ at rest, solve for the individual Nernst potentials for each ionic species, and also, from (3.2.18), for the resting potential. [Note that only permeability ratios are needed for this, but permeabilities change with the voltage, so we are using their values at resting potential to determine the voltage at rest.]

Note that the resting potential determined by (3.2.18) will not, in general, coincide with any of the Nernst potentials for individual ions given by (3.2.16). Thus, one would expect passive ionic flow to occur, thus changing the concentration ratios. This is where the active ionic pumps, referred to above, enter: they balance individual passive ionic flows, and so can maintain concentration ratios at equilibrium.

We now further explore the relationships between membrane currents and potentials and electrical circuits, and introduce a second and much simpler model for ionic currents. Assuming that the membrane has constant capacitance per unit area, the cross-membrane ion current is proportional to the rate of change of membrane potential, with the capacitance the constant of proportionality. There is also a negative sign, since a positive current (outward), results in reduced potential inside the cell:

$$C_m \frac{dV}{dt} = -I_{ion}. \quad (3.2.19)$$

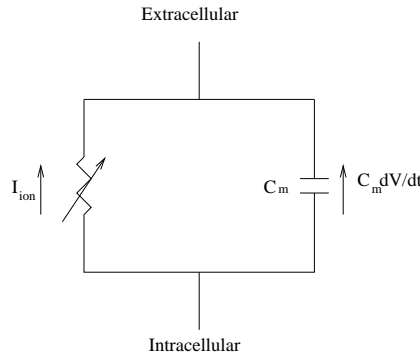


Figure 45: The cell membrane as an electrical circuit, with ionic and capacitive currents labeled as in Equation 3.2.19.

Equation (3.2.19) can be understood from Kirchhoff's loop law applied to the circuit of Fig. 45. The ionic current depends on the membrane conductance for that ion, as well as the Nernst and membrane potentials. Each ion has its own Nernst potential, and since different channels are permeable to specific ions, each ion has a different membrane conductance, which usually depends on the membrane potential. Various conductance forms are used in the H-H model. Our second model for ionic currents assumes a linear $I-V$ relationship. Denoting by V the membrane potential, g_S the specific ion's membrane conductance, and V_S its Nernst potential, the following simple equation replaces (3.2.15) for the ionic current:

$$I_S = g_S(V - V_S). \quad (3.2.20)$$

Constant conductance g_S implies a linear or *ohmic* relationship between potential difference from Nernst potential and current (Eqn. (3.2.20) is Ohm's law). However, cells usually exhibit voltage-dependent membrane conductances, and the voltage-current relationship (3.2.20) becomes nonlinear, as it will in the H-H equations. Recall that what is relevant here is the difference $(V - V_S)$ from *equilibrium* Nernst potential, *not* the potential difference from zero, and that this equation holds regardless of whether the ion is positively or negatively charged. This is because if $(V - V_S) < 0$, positive ions will flow in and negative ions will flow out, both of which are *negative* currents. Similarly, for $(V - V_S) > 0$, positive ions flow out, and negative ions flow in, both of which are *positive* currents.

We can use the ohmic conductance model to solve for the resting potential of the squid giant axon. With V_S denoting the Nernst potential of ion S , and including the three species K^+ , Na^+ , Cl^- , we have:

$$C_m \frac{dV}{dt} = -g_K(V - V_K) - g_{Na}(V - V_{Na}) - g_{Cl}(V - V_{Cl}), \quad (3.2.21)$$

which yields the resting potential upon setting the right hand side equal to zero. Note that the sign of each current changes when V passes through V_S ; hence the Nernst potential V_S is sometimes called the *reversal potential* for the ion S .

Exercise 13. *Using the Nernst potentials calculated in Exercise 12, and assuming the same ratios of conductances as permeabilities, solve (3.2.21) for the resting potential V and compare to the result of Exercise 12.*

Substantially more information on these and related topics can be found in [124, Chaps. 3-4] and [120, Chaps. 2-3]. The latter reference covers Fick's law, with clear derivations of the Einstein, Nernst-Planck, and Nernst equations. Section 2.2.4 has a good explanation of space-charge neutrality, which is either ignored or confusing in other references. The GHK equations are clearly derived, and there are clear examples of how to use the concepts, including a well-worked out resting potential exercise on p. 30. Chapter 3 has equivalent circuit descriptions.

3.3 The Hodgkin-Huxley (H-H) equations

The Hodgkin-Huxley (H-H) model of the squid giant axon [104] is a triumph of mathematical modeling. Their formulation of an ODE model that reproduces the action potential, and extension to a partial differential equation (PDE) to explain its propagation along the axon, marks the beginning of mathematical neuroscience. The model was built on 10+ years of beautiful and painstaking experimental work, reported in an impressive series of papers [105, 106, 101, 102, 103] that eventually gained Hodgkin and Huxley a Nobel prize. For an introduction with some historical notes, see [124, §4.1]. Here we give only the essence of the final mathematical model from [104].

Exercise 14. *A possible journal club team project: Read and present the experimental H-H(-K) papers [106, 101, 102, 103].*

The H-H model comprises four coupled nonlinear first-order differential equations, listed below,

and followed by notes on their derivation, explanations and analysis:

$$C_m \frac{dv}{dt} = -\bar{g}_K n^4 (v - v_K) - \bar{g}_{Na} m^3 h (v - v_{Na}) - g_L (v - v_L) + I_{app}, \quad (3.3.1a)$$

$$\frac{dm}{dt} = \alpha_m (1 - m) - \beta_m m, \quad (3.3.1b)$$

$$\frac{dn}{dt} = \alpha_n (1 - n) - \beta_n n, \quad (3.3.1c)$$

$$\frac{dh}{dt} = \alpha_h (1 - h) - \beta_h h. \quad (3.3.1d)$$

Equation (3.3.1a) expresses the change in membrane potential in terms of four currents, the potassium ion current, the sodium ion current, a leakage ion current, and an external applied current, according to Kirchhoff's law applied to the equivalent circuit of Fig. 46. The simple circuit of Fig. 45 lumps all the currents; here they are distinguished by the resistors in parallel. Bars have been added to the leading terms in the sodium and potassium conductances to indicate that these are now constant parameter values that multiply time-dependent functions $n(t)$, $m(t)$ and $h(t)$ to form the “dynamical” conductances $g_K = \bar{g}_K n^4$ and $g_{Na} = \bar{g}_{Na} m^3 h$ of Eqn. (3.2.21). The leak conductance $g_L = g_{Cl}$, primarily due to chloride ions, remains constant.

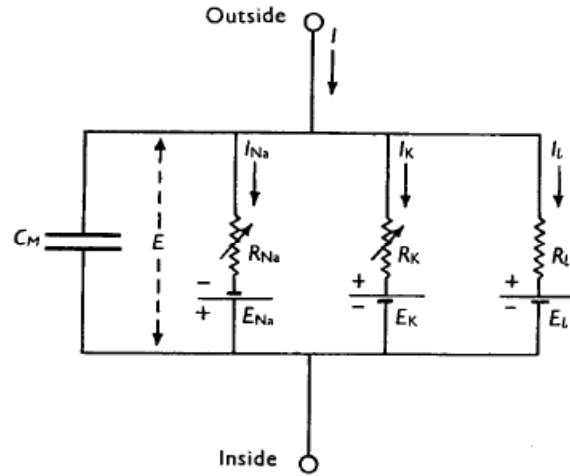


Figure 46: The equivalent circuit for the giant axon of squid, reproduced from [104, Fig. 1]. The leak conductance g_L is assumed constant, but the sodium and potassium conductances vary, indicated by the arrow crossing the resistor symbol. Batteries in series with each resistor represent the reversal potentials. Here the external current is labeled I .

The external current I_{app} may derive from synaptic inputs, electrical (gap junction) contacts with other cells, or from an intracellular electrode (it is sometimes called I_{ext} in that case). The sodium and potassium conductances are complicated and are described in terms of constants \bar{g}_{Na} and \bar{g}_K multiplied by functions of *channel gating variables* n , m and h , whose evolutions are described by the three succeeding equations, whose coefficients depend on voltage as described below. The leakage current $g_L(v - v_L)$ combines the effect of all other ions, the most important of which is chloride, and its membrane conductance g_L is assumed to remain constant over the relevant voltage range. Voltage dependences of the ion channels are also characterized by their *reversal potentials* v_K , v_{Na} , v_L : as the name suggests, the directions of the currents change as the membrane potential crosses these values.

A Note on units: Units are not always used in a consistent manner. So that numerical values remain reasonable, however, the following choices are often made. Membrane voltage: milliVolts (mV); currents: microAmps (μA); capacitance: microFarads (μF); conductance: milliSiemens = $10^{-6} \times 1/\text{milliOhm}$ (mS). The units of Eqn. (3.3.1a) are then μA for the cell as a whole. Cell size can be removed by expressing membrane capacitance and ion channel conductances as $\mu\text{F}/\text{cm}^2$ and mS/cm^2 , giving current densities per unit area in $\mu\text{A}/\text{cm}^2$. The gating variables m , n and h are dimensionless and take values in the unit interval $[0, 1]$. The usual time scale is milliseconds (ms).

In (3.3.1b-3.3.1d), each of the α 's and β 's is a function of voltage empirically fit by Hodgkin and Huxley to voltage clamp data (see below). The resulting six functions have units of ms^{-1} and depend on voltages in mV measured with respect to the resting potential of the cell. The constants are $g_{Na} = 120$, $g_K = 36$, and $g_L = 0.3$. $v_{Na} = 115$, $v_K = -12$, and $v_L = 10.6$. Note that the resting potential is very close to the potassium Nernst potential, compared to the sodium potential. This is because the equilibrium conductance is higher for potassium than for sodium, and it can drive the resting potential closer to the potassium Nernst potential. The forms of the α and β functions are:

$$\alpha_m = 0.1 \frac{25 - v}{\exp\left(\frac{25-v}{10}\right) - 1}, \quad \beta_m = 4 \exp\left(\frac{-v}{18}\right), \quad (3.3.2)$$

$$\alpha_h = 0.07 \exp\left(\frac{-v}{20}\right), \quad \beta_h = \frac{1}{\exp\left(\frac{30-v}{10}\right) + 1}, \quad (3.3.3)$$

$$\alpha_n = 0.01 \frac{10 - v}{\exp\left(\frac{10-v}{10}\right) - 1}, \quad \beta_n = 0.125 \exp\left(\frac{-v}{80}\right). \quad (3.3.4)$$

Sometimes the gating equations are rewritten in the form

$$\tau_n(v) \frac{dn}{dt} = n_\infty(v) - n, \quad \text{where } n_\infty(v) = \frac{\alpha_n(v)}{\alpha_n(v) + \beta_n(v)} \quad \text{and} \quad \tau_n(v) = \frac{1}{\alpha_n(v) + \beta_n(v)} \quad (3.3.5)$$

(and similarly for m and h), to emphasize the equilibrium potential $n_\infty(v)$ for the gating variable, and the time scale $\tau_n(v)$.

As noted above, the H-H model was developed through ingenious experimental work by Hodgkin, Huxley, Katz, and others. They used a *voltage clamp* technique that built on earlier work by Cole and Marmont. A metallic conductor was threaded inside the axon along its full length, thus inducing a uniform membrane potential so that voltage is independent of space. Voltage can then be stepped from one value to another and held constant, and the current $I_{app}(t)$ required to maintain it recorded as a function of time. Since there is no accumulation of charge in the axon, the applied current exactly balances the transmembrane ionic currents. Combined with manipulations of the extracellular solution (replacing sodium with inert choline, for example), this device permitted characterizations of the individual ionic currents, leading to the gating functions of (3.3.2-3.3.4) and the polynomial fits n^4 and m^3h for conductance dependence in (3.3.1a). Keener and Sneyd [124, §4.1] provide a good summary, which we now sketch.

In their experiments, Hodgkin and Huxley noticed that during a spike an initial inward current was followed by outward current flow. They hypothesized that the inward current was due to sodium ions, since open sodium channels would allow the higher extracellular concentrations to flow inward. The outward current was assumed to be due to potassium ions, given the higher potassium

concentrations inside the cell. (Recall that a hyperpolarizing current makes the membrane potential more negative, and a depolarizing current makes it more positive.) Separating the potassium current as described above, its conductance was seen from experimental data to exhibit a sigmoidal increase and an exponential decrease. Hodgkin and Huxley modeled this in terms of the fourth power of a new variable, $n(t)$: the potassium activation. The fourth power was not physiologically motivated, but was the lowest exponent that provided an acceptable fit to the data (a biophysical interpretation of this exponent will be given later). A mathematical examination of the response of n to a step increase and a step decrease of V , shows that $n(t)$ raised to the fourth power reproduces the sigmoidal-like increase from $n = 0$ and exponential decrease seen in experiments (cf. [104, Figs. 2 and 3]).

Exercise 15. Solve the ODE (3.3.1c) (or (3.3.5)) for $n(t)$ with an applied voltage step of magnitude V_a .

Exercise 16. Numerically integrate the ODE (3.3.1c) (or (3.3.5)) in response to a voltage step from 0, which had been a steady state, to 60 mV for 10 ms, followed by a voltage step back to 0 and let run until n settles to a steady state. Also, plot the exact solution from Exercise 15 for these same parameters; you will need piecewise solutions, since V_a will change at 10 ms.

The spike initiation dynamics suggested that sodium conductance is more complicated, involving *two* processes, one of which rapidly activates the sodium current, while the other turns it off more slowly [104, Fig. 6]). This led to proposal of sodium activation and deactivation variables $m(t)$ and $h(t)$. Acceptable fits were again obtained by a power law: m^3h . This elegant solution models the conductance by two underlying processes. Activation $m(t)$ starts small (near 0) and increases quickly, causing the initial inward sodium current, while inactivation $h(t)$ starts near 1 and slowly decreases, eventually shutting off the sodium current. The α and β functions are again chosen to match experimental data. Here, there is no obvious physiological explanation, but the goal of modeling is to use empirical data fits as little as possible, but to fill in the areas where the physiology is not understood. This allows the physiological model to have the correct interactions (matching data) with the parts of the system that are not physiologically modeled.

Fig. 47 shows the time course of a typical action potential simulated by numerical solution of the Hodgkin-Huxley equations (3.3.1), and Fig. 48 shows the time course of the gating variables over the same time period. Note the four phases of the action potential. First, there is a sharp increase in membrane potential called the upstroke as the sodium conductance increases quickly and drives the potential up towards the sodium Nernst potential. This excited or depolarized state does not last long. At these higher voltages, h decreases, lowering the sodium conductance, while n increases, increasing the potassium conductance and driving the voltage down towards the potassium potential. During the ensuing refractory period, no more action potentials are possible; m recovers quickly to its resting value, but n stays high and h remains low for a time, since their equations have longer time constants, thus holding the potential low and preventing it from spiking again. When n and h return to values that will allow another spike, the neuron is said to be in its recovery phase.

In interpreting the auxiliary variables m , n and h in terms of channels in the cell membrane that allow the ions to pass, we appeal to the statistics of large numbers. A given channel is open or closed, depending on the states of *gates* within it. Specifically, each gate within a channel is stochastically open or closed with a probability that depends on the voltage and the equation governing its *gating*

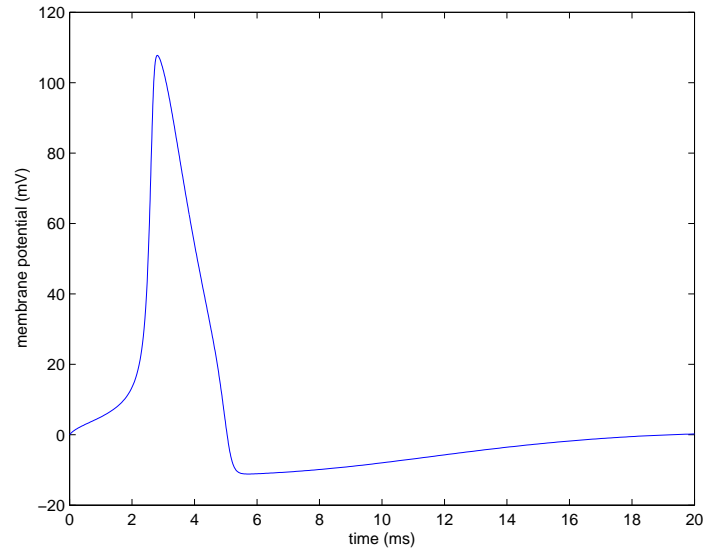


Figure 47: The time course of membrane voltage during an action potential and the subsequent refractory period. Note that the voltage scale has been shifted so that the resting potential is at 0 mV.

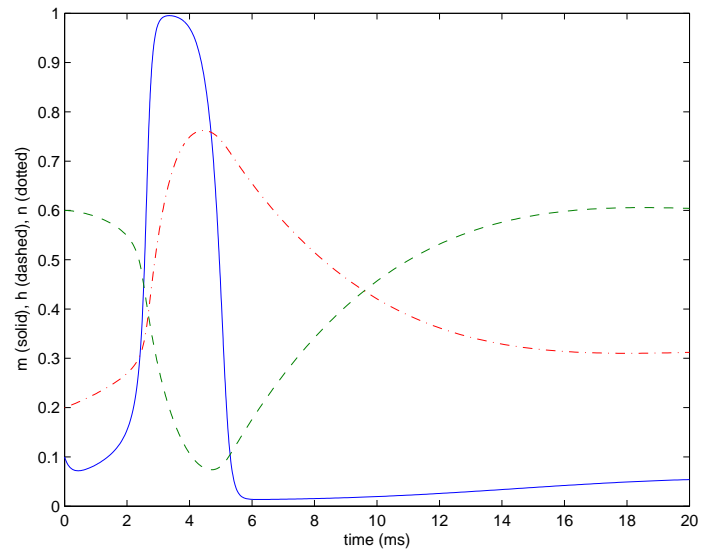


Figure 48: Gating variable evolutions during an action potential and the refractory period: m solid, n dash-dotted and h dashed. Note the differing timescales and the approximate anticorrelation of $n(t)$ and $h(t)$.

variable. Thus, n determines the probability that one of four gates in a potassium channel is open, so that the percentage of open channels is proportional to n^4 . Over a large number of ion channels, the probability that a given one is open is very close to the percentage that are actually open. The sodium channel is modeled as having three m gates and one h gate, all of which must be open for the channel to be open. Then the percentage of open gates is m^3h .

The numerical simulations of Figs. 47-48 give some information on the H-H equations, but it would be nice to have a more comprehensive picture. Such an analysis is necessarily more qualitative, because we are not interested in the numerical results in a single case, but rather in the general solution patterns for a range of parameters and initial conditions. In the next section we sketch two-dimensional simplifications of H-H that reveal the mathematical mechanisms for spiking more clearly.

In addition to the H-H papers cited at the beginning of this section, good supplements and sources for the material include Keener and Sneyd [124, Chap 5] and Johnston and Wu [120]. The latter has general information on membrane conductance models in Chapter 5, concluding with a gate model leading in to Chapter 6 on the H-H equations. Chapter 7 contains a detailed discussion of different potassium, sodium, calcium, and other ionic currents, including high-threshold calcium, low-threshold calcium, calcium-gated potassium, etc... Chapters 8-10 provide good coverage of stochastic models of ion channels. Chapter 8 is an introduction to the molecular structure of ion channels, Chapter 9 is an introduction to basic probability, and Chapter 10 uses the Chapman-Kolmogorov equation to analyze transition schemes for channel gating. These topics are very interesting but go considerably beyond the detail needed for this course. A final project could certainly be done in this area.

3.4 Two-dimensional simplifications of HH

In this section, we introduce two simplifications of the Hodgkin-Huxley equations, a two-dimensional reduction of HH, and the FitzHugh-Nagumo (F-N) equations. See [231, Chaps 8-9]. Although the FitzHugh-Nagumo model predated the HH reduction, we start with the latter. It was first proposed by Krinsky and Kokoz [131] and subsequently and independently, by Rinzel [180]. We follow Rinzel's description.

Exercise 17. *Plot projections of solutions of the H-H equations (3.3.1) in a repetitively spiking mode with input current $I = 15$, onto planes defined by the following pairs of state variables: (n, h) , (m, n) , and V, n . Explore the parameter space a little. See Fig. 49 for an example.*

Examining the behavior of the four H-H state variables, Rinzel noted that $m(t)$ changes relatively rapidly because its timescale $\tau_m = 1/(\alpha_m + \beta_m)$ is small relative to τ_n and τ_h in the relevant voltage range. See Eqns. (3.3.2-3.3.5) and Figs. 48 and 49. He therefore ignored transients in m and assumed that it is always equilibrated so that $\dot{m} = 0$, implying that

$$m(t) = m_\infty(v) = \frac{\alpha_m(v)}{\alpha_m(v) + \beta_m(v)}, \quad (3.4.1)$$

from (3.3.2) (cf. Eqn. (3.3.5)). He further noted that the variables $n(t)$ and $h(t)$ are approximately anti-correlated in the sense that throughout the action potential and recovery phase they remain

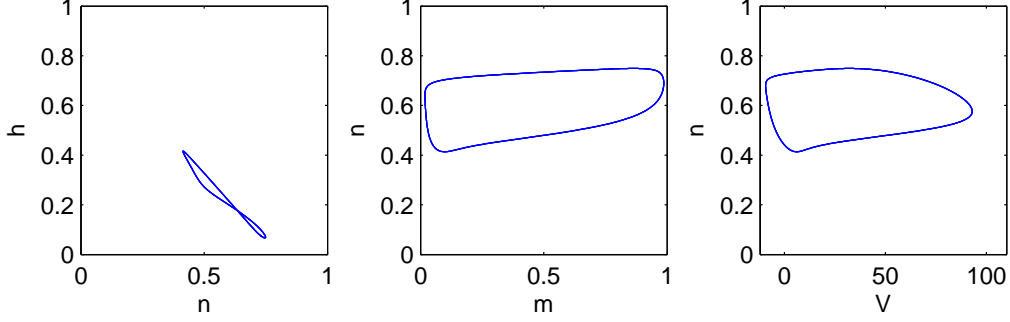


Figure 49: Phase plane projections for Eqns. (3.3.1). (Left) h plotted versus n : note that $h \approx 0.8 - n$. (Center) n plotted versus m : note that n changes little during the rapid change in m , and that m is approximately constant during the slow change in n . (Right) n plotted versus V : these are the two variables retained in Rinzel’s reduction.

close to a line of slope -1 : $h = a - n$ (Figs. 48 and 49). This allowed him to eliminate m and h as state variables, dropping Eqns. (3.3.1b) (3.3.1d) and replacing m and h in (3.3.1a) by the function $m_\infty(v)$ and by $h = a - n$, and reducing the system to two variables:

$$C_m \frac{dv}{dt} = -g_K n^4 (v - v_K) - g_{Na} m_\infty(v)^3 (1 - n)(v - v_{Na}) - g_L (v - v_L) + I_{app}, \quad (3.4.2a)$$

$$\frac{dn}{dt} = \alpha_n(v) (1 - n) - \beta_n(v) n. \quad (3.4.2b)$$

Note: Wilson [231, §9.1] gives the value $a = 1$, but for the “classical” parameters of the Hodgkin-Huxley paper [104], $a = 0.8$ is more appropriate: see Fig. 49.

We can now use nullclines and other techniques described in §2.4 to study the two-dimensional phase portrait of (3.4.2). However, while the $\dot{n} = 0$ nullcline

$$n = \frac{\alpha_n(v)}{\alpha_n(v) + \beta_n(v)} \quad (3.4.3)$$

can be written explicitly as a function of n in terms of V , the $\dot{v} = 0$ nullcline, given by setting the right-hand side of (3.4.2a) equal to zero, demands solution of a quartic polynomial in n . This can be done numerically to yield the phase portrait of Fig. 50. The lefthand plot is for $I_{app} = 0$ and features an attracting fixed point near $v = 0$ and two other fixed points (a saddle and a source). The righthand plot, for $I_{app} = 15$, shows a limit cycle which corresponds to periodic spiking. The plot of voltage versus time is seen in the left panel of Fig. 51. The spikes are qualitatively similar to those of Fig. 47 produced by Equation (3.3.1). Finally, in the right panel of Fig. 51, we see the (v, n) -phase plane. With no applied current, states settle to the stable fixed point at $V = 0$. With $I_{app} = 15$, however, the states approach a stable limit cycle, representing to periodic spiking behavior. The horizontal parts of the limit cycle correspond to fast flows in the phase plane seen in Fig. 50.

This figure also illustrates the threshold potential v_{th} , at the local minimum of the $\dot{v} = 0$ nullcline. When the fixed point lies to the left of this, as it does for $I_{app} = 0$, solutions may spike in response to perturbations that push v past v_{th} , but absent further perturbations the state will settle at the stable fixed point. When it moves to the right of v_{th} ($I_{app} = 15$) it loses stability and solutions repeatedly reach and cross threshold, leading to autonomous spiking.

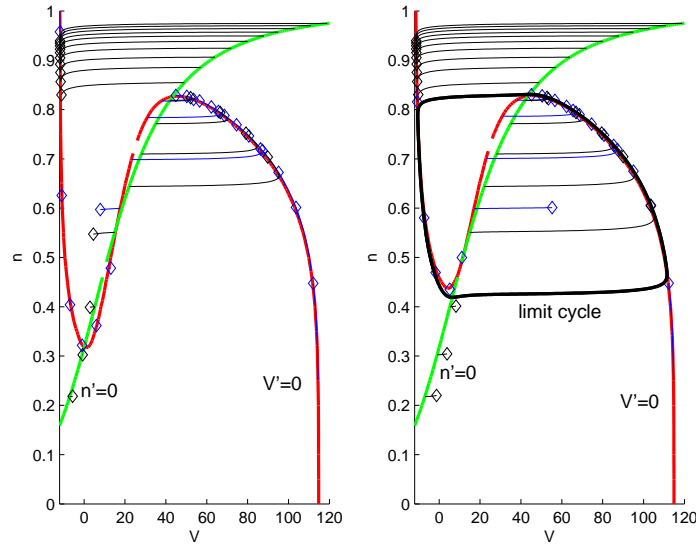


Figure 50: Phase planes of the reduced H-H equations (3.4.2), showing nullclines (bold) for $I_{app} = 0$ (a) and $I_{app} = 15$ (b). Diamonds are at the end of flows and thereby indicate the direction of the vector field. Approximately horizontal components correspond to fast flows and solutions move slowly near $\dot{v} = 0$: the slow manifold.

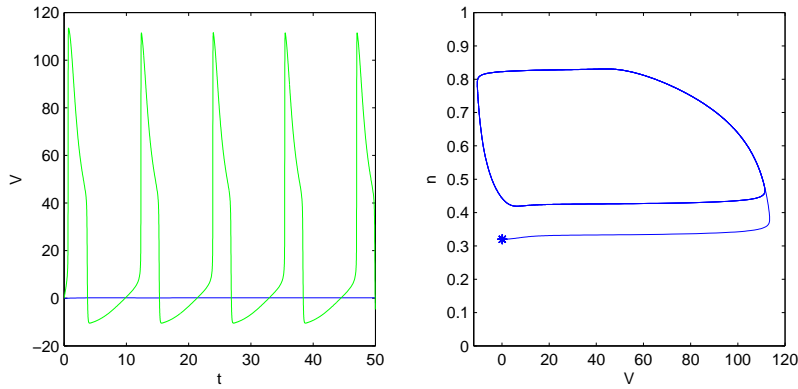


Figure 51: a) Plot of voltage V versus time t for (3.4.2) with zero applied current (flat) and for $I_{app} = 15$ (spikes). Note that spike shape is similar to that of full H-H, Fig. 47. b) With $I_{app} = 0$ the (V, n) phase plane contains a stable fixed point at $V = 0$, marked *, but with $I_{app} = 15$, the state leaves that point and approaches a limit cycle corresponding to spiking behavior.

The F-N equations have a similar *mathematical* structure to that of Rinzel's H-H reduction in that the nullclines have similar qualitative forms, but they are defined by simple cubic and linear functions rather than the complicated exponential sigmoids of (3.3.2-3.3.4). The physiological interpretation and physical units apart from the two timescales have vanished, but the qualitative properties remain. We reproduce the version from [231, §8.3]:

$$\dot{v} = \frac{1}{\tau_v} \left(v - \frac{v^3}{3} - r + I_{app} \right), \quad (3.4.4a)$$

$$\dot{r} = \frac{1}{\tau_r} (-r + 1.25v + 1.5). \quad (3.4.4b)$$

Wilson chooses values $\tau_v = 0.1$ and $\tau_r = 1.25$, which we adopt for the calculations to follow, but we write the time constants as parameters to indicate that they may vary (depending on temperature, for example). Here v still represents the membrane voltage, but r is a combined effective gating/recovery variable. The time constants $1/10$ (fast) and $1/0.8$ (slow) have been chosen to reflect the rapid upstroke and downstroke in the action potential (cf. Fig. 47) and the slower hyperpolarized, subthreshold dynamics, but as one can see from the time courses plotted in Fig. 52, the relative durations of the depolarized and hyperpolarized episodes are approximately equal, unlike the H-H dynamics of Figs. 47-48. The reason for this becomes clear when we examine the nullclines

$$\dot{r} = 0 : r = 1.5 + 1.25v \quad \text{and} \quad \dot{v} = 0 : r = v - \frac{v^3}{3} + I_{app} \quad (3.4.5)$$

shown in the phase portrait of Fig. 53. Provided that $\tau_v \ll \tau_r$ (and $0.1 \ll 1.25$ in this case), the vectorfield of (3.4.4) is dominated by its large \dot{v} component everywhere except in an $\mathcal{O}(|1/\tau_v|)$ neighborhood of the $\dot{v} = 0$ nullcline. The flow therefore moves approximately horizontally and quickly towards this *slow manifold* and follows it closely in the direction determined by the slower component (3.4.4b) of the vectorfield. This leads to the slow climb up the righthand branch of the cubic $\dot{v} = 0$ nullcline and the slow descent of its left-hand branch, punctuated by fast jumps up and down in voltage when the solutions leave the neighborhood of the attracting branches of the nullcline.

Note that (3.4.5) has a single equilibrium point for all values of I_{app} , as one can see by examining the cubic equation that results from setting $r = 1.5 + 1.25v$ (ensuring $\dot{r} = 0$ in $\dot{v} = 0$ or $r = v - \frac{v^3}{3} + I_{app}$, namely:

$$\frac{v^3}{3} + 0.25v + 1.5 - I_{app} = 0. \quad (3.4.6)$$

Alternatively, note that the cubic $\dot{v} = 0$ nullcline of (3.4.5) satisfies $\frac{dr}{dv} = 1 - v^2$ and so has maximum (positive) slope 1, while the linear nullcline $\dot{r} = 0$ has slope 1.25. Hence they can only intersect once. However, as I_{app} varies, the stability of this equilibrium can change.

Exercise 18. Show that, as I_{app} increases in (3.4.4), a Hopf bifurcation takes place, leading to the birth of a stable limit cycle. See [231, §8.3].

Exercise 19. Compute and compare bifurcation diagrams of full H-H equations (3.3.1) with those for the reduced H-H system (3.4.2) and for the F-N equations (3.4.4). Use AUTO and get help in the computer lab section of the course.

Wilson [231, §9.2] presents a simplification of of Rinzel's reduction (3.4.2) that also uses cubic and linear functions, but which better captures the spike/recovery dynamics and the spike rate vs. applied current characteristic.

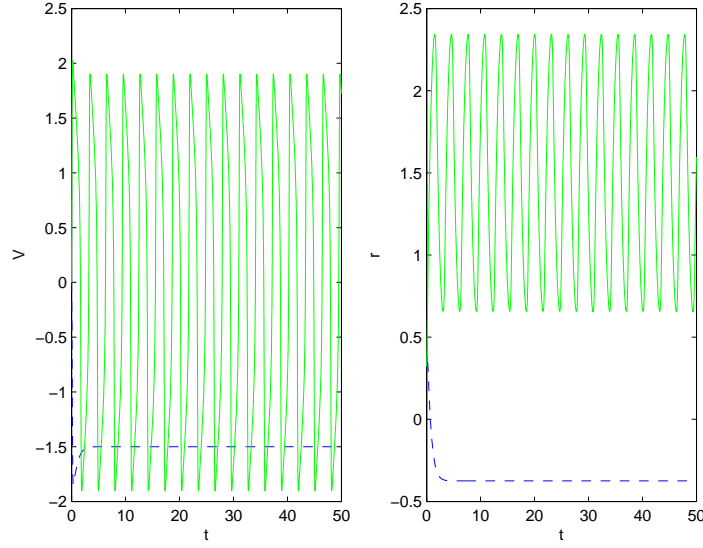


Figure 52: The time courses of the voltage and recovery variables v (left) and r (right) of the FitzHugh-Nagumo equations (3.4.4) for $I_{app} = 0$ (blue dotted) and $I_{app} = 1.5$ (green solid).

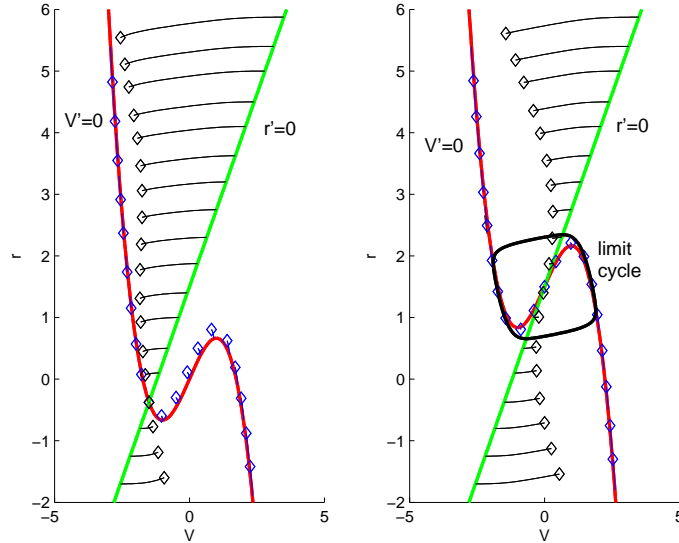


Figure 53: The phase plane of the FitzHugh-Nagumo equations (3.4.4), showing nullclines and indicating fast and slow flows for $I_{app} = 0$ (left) and $I_{app} = 1.5$ (right). On the left, there is an attracting fixed point, but on the right there is a limit cycle corresponding to the spiking seen in Fig. 52. Solutions shown as in Fig. 50.

3.5 Bursting neurons

Thus far we have considered models such as that of H-H that produce single spikes driven by fast depolarizing (Na^+) currents, separated by refractory periods controlled by slower hyperpolarizing (K^+) currents. Bursting – the clustering of multiple spikes followed by a refractory period of relative quiescence – can also occur and can vary substantially in form and function [124, Chap. 6] and [231, Chap. 10]. The mechanism can be described qualitatively as the interaction of two subsystems dynamically separated by their intrinsic time scales: a faster one, typically governed by the sodium and potassium channels, which can either be at rest or exhibit (periodic) oscillations, and a slow subsystem driving the first through its quiescent and oscillatory states in a quasi-static manner. The slower mechanism can be attributed to accumulation of intracellular calcium ions (referred to as calcium dynamics [124]) that mediates a hyperpolarizing potassium current, or to other slow voltage-dependent processes.

Singular perturbative reduction methods [121], mentioned in passing in §3.6.1, can also provide insights into the dynamics of bursting neurons, which may be written in the general form

$$\dot{\mathbf{u}} = \mathbf{f}(\mathbf{u}, c), \quad (3.5.1a)$$

$$\dot{c} = \epsilon g(\mathbf{u}, c), \quad (3.5.1b)$$

where the vector $\mathbf{u} = (v, \mathbf{w}) \in \mathbb{R}^n$, v denotes the cell membrane voltage, $\mathbf{w} = (w_1, \dots, w_{n-1})$ represents a collection of $n - 1$ gating variables w_i , and $\epsilon \ll 1$ is a small parameter. The variable c may represent calcium concentration or, more generally, any (very) slowly varying quantity responsible for bursting.

The subset of *fast* equations (3.5.1a) generally takes the Hodgkin-Huxley (HH) form (3.3.1):

$$C\dot{v} = -I_{\text{ion}}(v, w_1, \dots, w_n, c) + I_{\text{ext}}(t) \quad (3.5.2a)$$

$$\dot{w}_i = \frac{w_{i\infty}(v) - w_i}{\tau_i(v)}; \quad i = 1, \dots, n - 1, \quad (3.5.2b)$$

where the term $I_{\text{ion}}(\dots)$ represents the sum of all ionic currents, and the functions $w_{i\infty}(v)$ and $\tau_i(v)$ take forms similar to those of the H-H system, cf. Eqns. (3.3.2-3.3.5). Some of the gating equations (3.5.2b) may be so much faster than others that those variables may be assumed to be equilibrated, but it is implicit that *all* are fast in comparison to the slow variable c .

Wilson [231, Chap. 10] and Keener and Sneyd [124, Chap. 9] both provide reviews of bursting models, including those proposed to describe oscillations in β -cells of the pancreas, where insulin is secreted. Here we merely sketch some of the key ideas, basing our discussion on the Sherman-Rinzel-Keizer (SRK) model [197] for pancreatic cells. This is a minimal burster model consisting of three ODEs, two fast and one slow:

$$C\dot{v} = -\bar{g}_K n(v - E_K) - I_{\text{Ca}}(v) - g_{\text{KCa}}(Ca) \cdot (v - E_K) \quad (3.5.3a)$$

$$\dot{n} = \delta \frac{n_{\infty}(v) - n}{\tau_n(v)} \quad (3.5.3b)$$

$$C\dot{a} = \epsilon(-\alpha I_{\text{Ca}}(v) - k_{\text{Ca}} Ca). \quad (3.5.3c)$$

where

$$I_{\text{Ca}}(v) = \bar{g}_{\text{Ca}} m_{\infty}(v) h_{\infty}(v) (v - E_{\text{Ca}}), \quad g_{\text{KCa}}(Ca) = \bar{g}_{\text{KCa}} \frac{Ca}{K_d + Ca}, \quad (3.5.4)$$

\bar{g}_{Ca}	$= 1400 \text{ pS}$	E_{Ca}	$= 110 \text{ mV}$
\bar{g}_K	$= 2500 \text{ pS}$	E_K	$= -75 \text{ mV}$
\bar{g}_{KCa}	$= 30000 \text{ pS}$		
C	$= 6310 \text{ fF}$	V_{Cell}	$= 1150 \text{ } \mu\text{m}^3$
F	$= 96.487 \text{ Coul/mMol}$	K_d	$= 100 \text{ } \mu\text{Mol}$
δ	$= 1.6$	k_{Ca}	$= 0.03 \text{ m/s}$
V_n	$= -15 \text{ mV}$	S_n	$= 5.6 \text{ mV}$
V_m	$= 4 \text{ mV}$	S_m	$= 14 \text{ mV}$
V_h	$= -10 \text{ mV}$	S_h	$= 10 \text{ mV}$
a	$= 65 \text{ mV}$	b	$= 20 \text{ mV}$
γ	$= 60 \text{ ms}$	\bar{V}	$= -75 \text{ mV}$
ϵ	$= 0.001$		

Table 1: Parameter values for the SRK model for bursting pancreatic β -cells.

and $n_\infty, m_\infty, h_\infty$ are standard HH-type equilibrium functions:

$$n_\infty(v) = \frac{1}{1 + e^{\frac{V_n - v}{S_n}}}, \quad m_\infty(v) = \frac{1}{1 + e^{\frac{V_m - v}{S_m}}}, \quad h_\infty(v) = \frac{1}{1 + e^{\frac{v - V_h}{S_h}}} \quad (3.5.5)$$

and the the n -timescale and parameter α are

$$\tau_n(v) = \frac{\gamma}{e^{\frac{v - \bar{V}}{a}} - e^{-\frac{v - \bar{V}}{b}}}, \quad \alpha = \frac{1}{2V_{Cell}F}. \quad (3.5.6)$$

The model has a potassium current $I_K = \bar{g}_K n(v - E_K)$, a fast transitory calcium current I_{Ca} and a very slow calcium-dependent potassium current $I_{KCa} = g_{KCa}(Ca) \cdot (v - E_K)$. Intracellular calcium Ca affects the conductance via (3.5.4) and has its own dynamics as given in the last equation of (3.5.3c). This system differs from the version in [124, p.192] in that the gating variable n appears linearly in (3.5.3a) rather than as n^4 . Specific parameter values are given in Table 1 and our presentation is taken from [77].

The calcium concentration Ca enters equation (3.5.3a) via the Hill-type function $g_{KCa}(Ca)$ of (3.5.4), and the fast system becomes simpler if one defines a new slow variable c in place of Ca , so that the calcium-dependent potassium current $I_{KCa} = g_{KCa}(Ca)(v - E_K)$ is linear in c :

$$c = \frac{Ca}{K_d + Ca}. \quad (3.5.7)$$

(In fact, since $K_d \gg Ca$, the conductance g_{KCa} is essentially proportional to c and (3.5.7) is in its linear régime.) Differentiating (3.5.7) we find $\dot{c} = \frac{K_d}{(K_d + Ca)^2} \dot{Ca}$, and inverting (3.5.7) to obtain $Ca = K_d \frac{c}{1 - c}$, the ODEs (3.5.3) become:

$$\begin{aligned} \dot{v} &= \frac{1}{C} [-\bar{g}_K n(v - E_K) - I_{Ca}(v) - \bar{g}_{KCa} c(v - E_K) + I_{ext}], \\ \dot{n} &= \delta \left[\frac{n_\infty(v) - n}{\tau_n(v)} \right], \\ \dot{c} &= \epsilon \frac{(1 - c)^2}{K_d} \left[-\alpha I_{Ca}(v) - k_{Ca} K_d \frac{c}{1 - c} \right], \end{aligned} \quad (3.5.8)$$

where we have added an external (bias) current I_{ext} and $I_{\text{Ca}}(v)$ is given in (3.5.4) above.

To understand the mathematical origin of bursting, first recall that we assume $\epsilon \ll \delta \ll 1/C$ (cf. the values in Table 1). Hence we can “freeze” c (or Ca) and consider the phase plane of the (v, n) subsystem for different values of c in its operating range. This subsystem is similar to the two-dimensional reductions of the H-H equation of §3.4, and it possesses a stable limit cycle (with relatively fast v and slower n dynamics) over a range of c values. The spiking frequency is basically set by the values of C and δ , which control the (relatively) slow flow along the $\dot{v} = 0$ nullcline and the fast jumps along $n = \text{constant}$ lines (cf. Figs. 50-51), but it can also vary with c . More significantly the limit cycle can disappear in local and global bifurcations as c varies. Bursting results from hysteretic transitions, driven by the slow variable c , between a quasi-static quiescent state of the fast (v, n) subsystem in which v remains subthreshold, and a periodic (spiking) state. Here c is regarded as a bifurcation parameter.

The branch of equilibria (or slow manifold) that forms the bifurcation diagram is obtained by setting $\dot{v} = \dot{n} = 0$, implying that $n = n_{\infty}(v)$ and $\bar{g}_K n_{\infty}(v)(v - E_K) + I_{\text{Ca}}(v) + \bar{g}_{\text{KCa}} c(v - E_K) = I_{\text{ext}}$, or

$$c = \frac{I_{\text{ext}} - \bar{g}_K n_{\infty}(v)(v - E_K) - I_{\text{Ca}}(v)}{\bar{g}_{\text{KCa}}(v - E_K)}. \quad (3.5.9)$$

This yields a cubic-like curve in the (c, v) plane with two folds. Laborious analyses, or numerical explorations, reveal that saddle-node bifurcations occur at the folds, and that a Hopf bifurcation occurs on the upper segment of the curve towards the left. See Fig. 54. The lower segment (to the right of the saddle-node) consists of stable sinks, the middle segment contains saddle points, and the upper one, between the Hopf bifurcation point and the saddle node, is filled with unstable sources.

To see how c evolves we examine the nullclines of the \dot{c} equation. We have $\dot{c} = 0$ on $c = 1$ and on

$$c = \frac{\alpha \bar{g}_{\text{Ca}} m_{\infty}(v) h_{\infty}(v)(v - E_K)}{\alpha \bar{g}_{\text{Ca}} m_{\infty}(v) h_{\infty}(v)(v - E_K) - k_{\text{Ca}} K_d}, \quad (3.5.10)$$

and away from $c = 1$ the sign of \dot{c} (and hence the direction in which c moves) is determined by the sign of the RHS of (3.5.10). The curve defined by (3.5.10) is also shown in Fig. 54: c moves to the right above and to the left of it, and moves to the left below it and to the right. Starting from initial data with the fast system state at a sink on the lower branch of Fig. 54, c will slowly decrease, while the fast state tracks the slow manifold of equilibria, until it arrives at the saddle node point, where the equilibrium disappears and the solution jumps rapidly up in v and converges on the limit cycle, leading to a burst of spikes. Since v lies above the $\dot{c} = 0$ nullcline in this regime, c now begins to increase, leading to a (slight) slowing of the spike rate, and ultimately “collision” of the limit cycle with the intermediate saddle point in a global, homoclinic bifurcation [89, §6.1] (to be described in class). Spiking terminates, the solution drops onto the lower segment of the slow manifold initiating the refractory period, and the process repeats.

As we have noted, the parameters C and δ primarily control spike rates within a burst, and the (very) small parameter ϵ clearly controls the rate of evolution of c , and hence the overall bursting period: $T_{\text{burst}} \sim 1/\epsilon$. The other parameters, including conductances and the external current, influence the behavior in more subtle ways. The *duty cycle* – the fraction of T_{burst} occupied by spiking – is controlled by the relative speeds with which c changes on the upper and lower segments

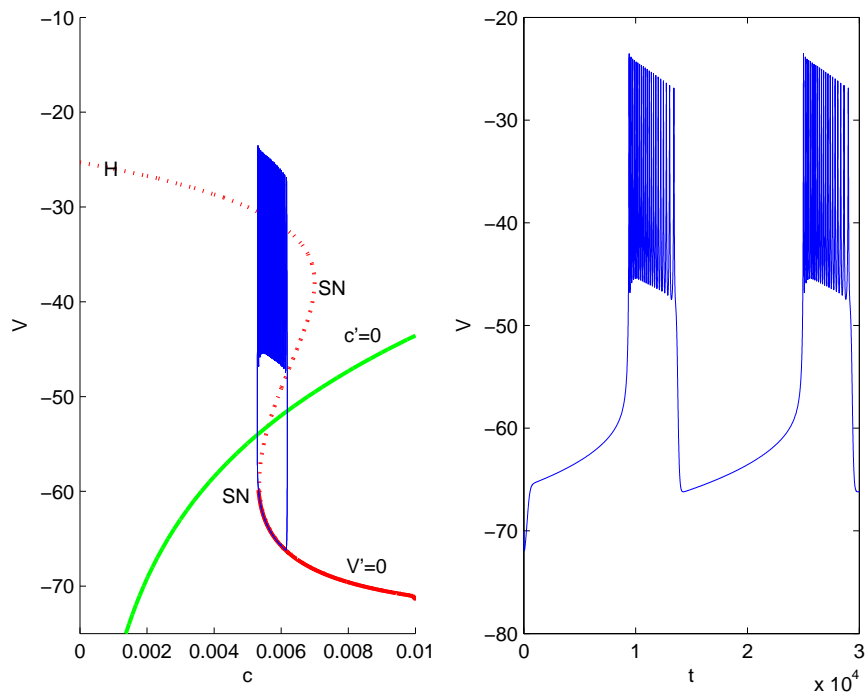
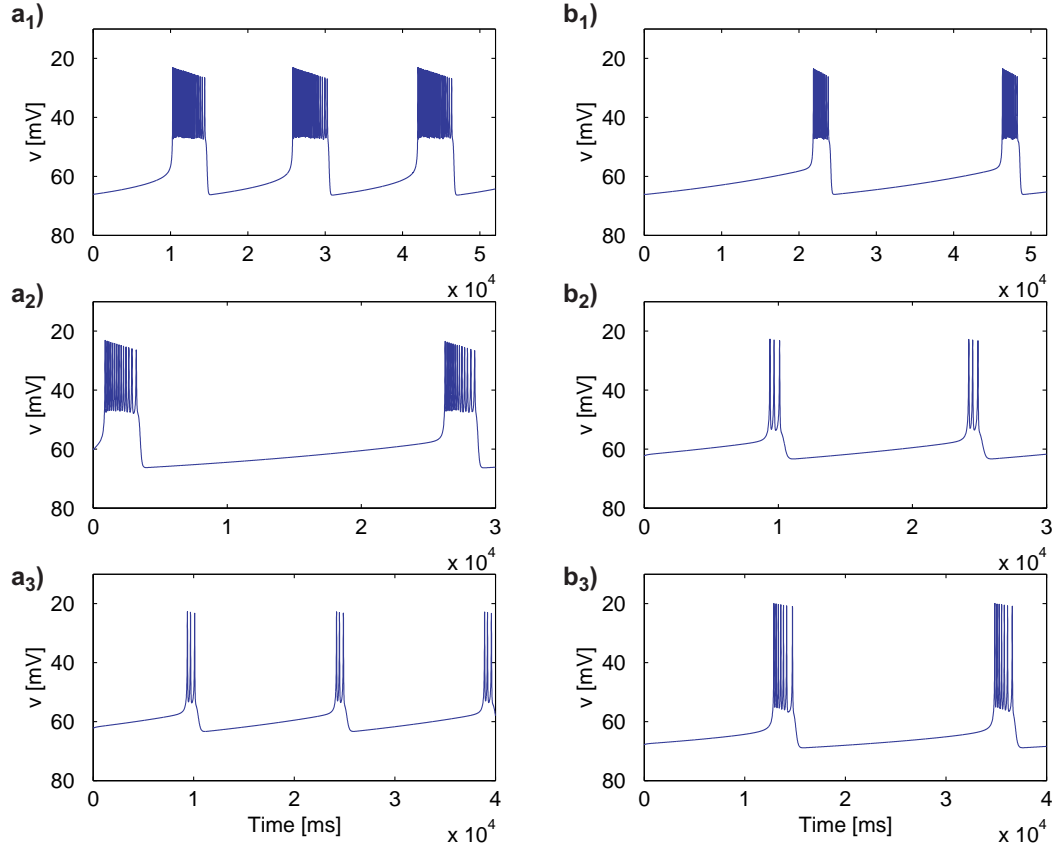


Figure 54: Left: A branch of equilibria for the “frozen c ” system and bifurcations with respect to c ; note the two saddle nodes and a Hopf bifurcation. The $\dot{v} = 0$ and $\dot{c} = 0$ nullclines and a typical bursting trajectory are projected onto the (c, v) plane. Right: The voltage time history exhibiting periodic bursts. Adapted from [77, Fig 11].



of the slow manifold, and this in turn is controlled by how close the $\dot{c} = 0$ nullcline lies to the those segments. From (3.5.9) we see that I_{ext} and the conductance \bar{g}_{KCa} affect the position of the latter, while $\alpha\bar{g}_{\text{Ca}}$ and $k_{\text{Ca}}K_d$ influence the $\dot{c} = 0$ nullcline (3.5.10). The effect of \bar{g}_{KCa} is simple: as the denominator in (3.5.9) it scales the horizontal extent of the slow manifold and its bifurcations in Fig. 54, so that increases in \bar{g}_{KCa} bring the homoclinic bifurcation leftward (and vice versa). If the $\dot{c} = 0$ nullcline is closer to the lower branch, this reduces the burst duration more than the refractory period, thereby decreasing the duty cycle (the bursting frequency also changes slightly). See Fig. 55(a₁-b₁).

Changes in membrane capacitance C have little effect on AP numbers (although they can drastically affect AP magnitudes), but decreasing the parameter δ reduces the number of APs from 22 to 2-3; this is accompanied by a moderate increase in bursting frequency: Fig. 55(a₂-b₂). The bias current I_{ext} also has a strong influence, permitting adjustment of AP numbers without drastically changing the bursting frequency: Fig. 55(a₃-b₃).

Note that if c (or Ca) is taken constant at a value in the range in which there are three equilibria in the bifurcation diagram of Fig. 54, then the fast equations (3.5.3a-3.5.3b) have coexisting stable equilibria (hyperpolarized states) *and* periodic spiking states. Such behavior is called *bistable*: the neuron can be “kicked” from inactivity to activity or vice versa by transient stimuli, thus providing short-term memory at the level of a single cell.

Both Wilson [231, Chap. 10] and Keener and Sneyd [124, Chap. 9] include dynamical taxonomies or classifications of different types of bursters based on the bifurcations that the fast \mathbf{u} subsystem undergoes as the slow c variable drifts back and forth, and [231, §10.1] contains a nice description of *spike frequency adaptation* due to a slow “afterhyperpolarizing” or I_{AHP} current. For (many) more details and analyses of general bursting models, also see [77].

3.6 Propagation of action potentials

Hodgkin and Huxley’s paper [104] also contains a model for the propagation of the AP along the axon, viewed as a one-dimensional continuum. The ODEs (3.3.1) and their simplifications considered in §§3.3-3.4 assume a spatially-uniform potential difference across the cell membrane⁸. While this may be reasonable for compact cells, or for the soma of a neuron, it is clearly wrong for cells with extensive axonal or dendritic processes and for rapid potential changes, and it cannot, of course, account for propagating APs that travel as waves along the axon and dendrites. A spatio-temporal description is needed.

The spatial description of current I_m passing through the cell membrane is given in terms of transmembrane potential v by the *cable equation*:

$$I_m = \frac{\partial}{\partial x} \left(\frac{1}{R} \frac{\partial v}{\partial x} \right). \quad (3.6.1)$$

(The diffusion or heat equation of §3.2 appears again!) Here $R = R_i + R_e$, which may depend on x , denotes the sum of the intra- and extracellular resistances per unit length (the intra- and

⁸As noted in §3.3, in some of their experiments Hodgkin and Huxley *enforced* spatial uniformity in the giant axon of the squid by threading a thin wire down its interior (voltage clamp).

extracellular media near the membrane carry the current along the cable). This equation may be derived as follows (cf. [124, Chap. 6]): first consider a short (infinitesimal) segment $[x, x + dx]$ of length dx , having intra- and extracellular resistences $R_i dx$, $R_e dx$, and denote the intra- and extracellular voltages at either end as $v_i(x)$, $v_e(x)$ and $v_i(x + dx)$, $v_e(x + dx)$: Fig. 56(a). With the convention of positive current flow from left to right (x to $x + dx$), Ohm's law implies that

$$v_i(x + dx) - v_i(x) = -I_i(x)R_i dx \quad \text{and} \quad v_e(x + dx) - v_e(x) = -I_e(x)R_e dx, \quad (3.6.2)$$

where I_i , I_e denote the internal and external currents in the axial direction. Dividing by dx and taking the limit $dx \rightarrow 0$, as in calculus, gives

$$I_i(x) = -\frac{1}{R_i} \frac{\partial v_i}{\partial x} \quad \text{and} \quad I_e(x) = -\frac{1}{R_e} \frac{\partial v_e}{\partial x}. \quad (3.6.3)$$

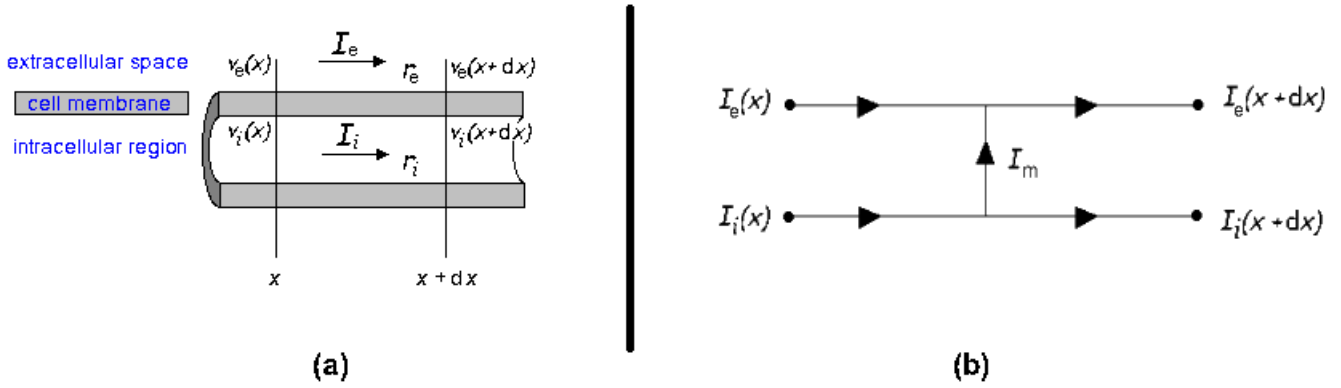


Figure 56: a) Current flow along an axon. b) Schematic for the Kirchhoff law current balance.

We next balance the axial currents I_i, I_e at x and $x + dx$ with the transmembrane current, appealing to Kirchhoff's laws. If I_m is the transmembrane current per unit cable length, this gives

$$I_e(x + dx) - I_e(x) = I_m dx = I_i(x) - I_i(x + dx) : \quad (3.6.4)$$

in words: what is lost in external current is gained in internal current. See Fig. 56(b) and recall that outward transmembrane currents are positive: §3.2. Dividing by dx and taking the limit again gives

$$I_m(x) = \frac{\partial I_e}{\partial x} = -\frac{\partial I_i}{\partial x}. \quad (3.6.5)$$

We now use the facts that $v = v_i - v_e$ and the total axial current $I_a = I_i + I_e$ is constant to eliminate explicit reference to the internal and external currents and voltages. From (3.6.3) and $v_e = v_i - v$ we get

$$\begin{aligned} I_a &= I_i + I_e = -\frac{1}{R_i} \frac{\partial v_i}{\partial x} - \frac{1}{R_e} \frac{\partial (v - v_i)}{\partial x} = -\left(\frac{R_i + R_e}{R_i R_e}\right) \frac{\partial v_i}{\partial x} + \frac{1}{R_e} \frac{\partial v}{\partial x} \\ \Rightarrow \frac{R_e I_a}{R_i + R_e} &= -\frac{1}{R_i} \frac{\partial v_i}{\partial x} + \left(\frac{1}{R_i + R_e}\right) \frac{\partial v}{\partial x}. \end{aligned} \quad (3.6.6)$$

Since $\partial I_a / \partial x = 0$ (I_a constant), this gives

$$\frac{\partial}{\partial x} \left[\frac{1}{R_i} \frac{\partial v_i}{\partial x} \right] = \frac{\partial}{\partial x} \left[\left(\frac{1}{R_i + R_e} \right) \frac{\partial v}{\partial x} \right] - I_a \frac{\partial}{\partial x} \left[\frac{R_e}{R_i + R_e} \right]. \quad (3.6.7)$$

If the resistances are constant, or if $R_e \approx 0$ (e.g. the axon is isolated in a large bath of fluid, as assumed in [104], cf. [124]) the final term is zero and substituting I_i for $\partial v_i / \partial x$ from the first of (3.6.3) and using (3.6.5) we obtain

$$I_m(x) = -\frac{\partial I_i}{\partial x} = \frac{\partial}{\partial x} \left[\left(\frac{1}{R_i + R_e} \right) \frac{\partial v}{\partial x} \right] = \frac{\partial}{\partial x} \left(\frac{1}{R} \frac{\partial v}{\partial x} \right). \quad (3.6.8)$$

To connect this with the H-H ODEs (3.3.1) we recall that the transmembrane current I_m is the sum of the capacitive current $C \frac{\partial v}{\partial t}$ and the ionic and applied currents, giving:

$$p \left(C \frac{\partial v}{\partial t} + I_{ion} - I_{app} \right) = \frac{\partial}{\partial x} \left(\frac{1}{R} \frac{\partial v}{\partial x} \right). \quad (3.6.9)$$

Here the LHS is multiplied by the perimeter p of the axon to put all terms into the same units: $\mu\text{Amps/cm}$ (recall that R_i and R_e are resistances per unit length of the cable and so have units mOhms/cm , the units of $1/R$ are $\text{mSeimens}\cdot\text{cm}$, and $\text{mSeimens}\cdot\text{cm} \times \text{mV/cm}^2 = \mu\text{Amps/cm}$). Partial derivatives are used here and above, because $v = v(x, t)$ and the currents also depend on t and x . Finally, note that the sign of I_{app} in (3.6.9) is chosen for consistency with (3.3.1), to which (3.6.9) reduces if v does not depend on x .

Like ODEs, PDEs such as (3.6.9) must be supplemented by initial conditions to be well-posed and admit unique solutions, but they also require *boundary conditions*, specifying the states which are assumed at the ends or edges of the spatial domain. Here we are considering one-dimensional cables (axons) and so we must specify the voltages at the end points. For endpoints $x = 0, L$ typical cases are:

$$v(0, t) = v_0, \quad v(L, t) = v_L \quad (\text{voltage clamp}), \quad (3.6.10)$$

$$\left. \frac{\partial v}{\partial x} \right|_{(0,t)} = 0, \quad \left. \frac{\partial v}{\partial x} \right|_{(L,t)} = 0 \quad (\text{insulated ends}), \quad (3.6.11)$$

$$\text{or} \quad \left. \frac{\partial v}{\partial x} \right|_{(0,t)} = -R_i I_i(0, t) \quad (\text{current injection at } x = 0); \quad (3.6.12)$$

and a suitable boundary condition is that the axon starts in its resting state

$$v(x, 0) = 0. \quad (3.6.13)$$

See [124, §§6.1-2] for more on the derivation of (3.6.9), and see Chap. 8 in general, and [120, Chap. 4] for additional material. Abbot and Dayan [50, §6.3] also derive (3.6.1), but not in detail. Wilson [231, Chap. 15] also has a brief derivation, and his chapter contains more on dendritic trees and spatially-discretized compartmental models of neurons.

3.6.1 Traveling waves in a simple PDE

Equation (3.6.9) is an example of a *reaction-diffusion equation* (RDE), the simplest types of which take the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(u), \quad (3.6.14)$$

where $f(u)$ describes the “reaction” or source term which depends on the state $u(x, t)$. Such equations describe the propagation of waves in active media, in contrast to the classical wave and heat equations

$$\frac{\partial^2 u}{\partial t^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial x^2} \quad \text{and} \quad \frac{\partial u}{\partial t} = \alpha^2 \frac{\partial^2 u}{\partial x^2} \quad (3.6.15)$$

which describe the evolution of disturbances in passive media.

Example 9. *Consider the bistable RDE*

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u - u^3, \quad (3.6.16)$$

so called because in the spatially-uniform case ($\partial^2 u / \partial x^2 \equiv 0$) (3.6.16) becomes the ODE $\dot{u} = u - u^3$ has two stable states $u^e = \pm 1$ separated by an unstable fixed point $u^e = 0$. These spatially uniform states $u^e(x) = \pm 1$ and $u^e(x) = 0$ are also equilibria for the RDE, but, as for nonlinear ODEs it is usually difficult to find time- (and space-) dependent solutions.

Here we can find space-dependent equilibria: solutions of the boundary value problem (BVP):

$$\frac{\partial^2 u}{\partial x^2} + u - u^3 = 0, \quad \left. \frac{\partial u}{\partial x} \right|_{(0)} = \left. \frac{\partial u}{\partial x} \right|_{(L)} = 0 \quad (3.6.17)$$

(we assume insulated boundaries). Indeed, writing $\frac{\partial u}{\partial x} = u'$ and $\frac{\partial^2 u}{\partial x^2} = u''$ and multiplying by u' , we may integrate (3.6.17) once:

$$\int u'' u' dx = - \int (u - u^3) u' dx \Rightarrow \frac{(u')^2}{2} + \frac{u^2}{2} - \frac{u^4}{4} = C, \quad (3.6.18)$$

where C is a constant of integration. From (3.6.18) we get

$$u' = \frac{du}{dx} = \sqrt{\left(C - \frac{u^2}{2} + \frac{u^4}{4}\right)} \Rightarrow \int \frac{du}{\sqrt{(C - u^2/2 + u^4/4)}} = \int dx = x - x_0 : \quad (3.6.19)$$

a separable first order ODE. A second constant of integration, x_0 has appeared, and it and C are implicitly determined by the boundary conditions $u'(0) = u'(L) = 0$ of (3.6.17). But alas, only for special values of C can the integral in (3.6.19) be evaluated (it is a Jacobian elliptic function – a special function – in general. However, we can deduce qualitative properties of solutions of the BVP by plotting level sets of its first integral (3.6.18): Fig. 57.

The closed curves surrounding the origin $(u, u') = (0, 0)$ for $0 < C < 1/4$ correspond to periodic orbits, whose periods increase monotonically from 2π to infinity as C goes from 0 to $1/4$ (the proof of this relies on properties of Jacobian elliptic integrals), and to satisfy the boundary conditions of (3.6.17) segments of these closed curves that start and end on $u' = 0$ must be selected. Potentially

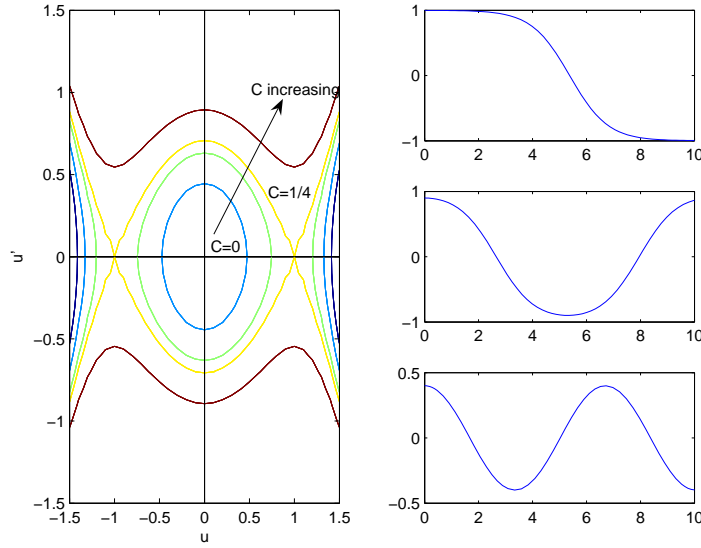


Figure 57: Level sets of (3.6.18) (a); solutions of the BVP (3.6.17) must begin and end on the horizontal axis $u' = 0$. The first three stationary solutions (b).

one can have $1/2, 1, 3/2$, etc. full turns, but these must be “fitted in” to the length L of the domain, and so **no** such solutions exist if $L < \pi$.

Time-dependent solutions are much harder, and here we only consider the special case of uniform traveling waves (TWs): spatial structures that move at constant velocity c and so may be written in the special form $u(x, t) = U(x - ct)$. Note that if $c > 0$ the wave travels in the direction of increasing x (from left to right). Letting $\zeta = x - ct$ represent the **wave-frame coordinate** and denoting derivatives with respect to ζ $dU/d\zeta = U'$, from the chain rule we have:

$$\frac{\partial U}{\partial x} = U' \frac{\partial(x - ct)}{\partial x} = U', \quad \frac{\partial U}{\partial t} = U' \frac{\partial(x - ct)}{\partial t} = -cU', \quad \text{and} \quad \frac{\partial^2 U}{\partial x^2} = U''. \quad (3.6.20)$$

Substituting these into (3.6.16) gives the ODE

$$-cU' = U'' + U - U^3 \Rightarrow U'' + cU' + U - U^3 = 0 : \quad (3.6.21)$$

an equation for a nonlinear oscillator with linear damping. The phase plane of (3.6.21) for $c > 0$ is shown in Fig. 58. We shall seek TWs defined on the real line $-\infty < \zeta < \infty$ (a reasonable simplification for cables (axons) that are long in comparison to their diameter), and require that $U'(x) \rightarrow 0$ as $x \rightarrow \pm\infty$, so that they connect uniform states. Fig. 58 illustrates that, for any $c > 0$ there are two special solutions, branches of the unstable manifolds of $(U, U') = (\pm 1, 0)$, that connect those fixed points to the sink at $(U, U') = (0, 0)$. These are called **heteroclinic orbits**, and they correspond to TWs in which U rises from -1 to 0 and drops from $+1$ to 0 respectively. If $c < 0$ the sink becomes a source and the connections flow, and waves propagate, in the opposite direction. Only for $c = 0$ do heteroclinic orbits connecting the equilibria $(U, U') = (\pm 1, 0)$ exist.

Exercise 20. Describe how the form of the heteroclinic traveling waves changes as wavespeed c increases for the bistable RDE in the example above. Show that there is a critical speed c_{mon} at and above which the wave profile is monotonic (unlike the oscillations in Fig. 58).

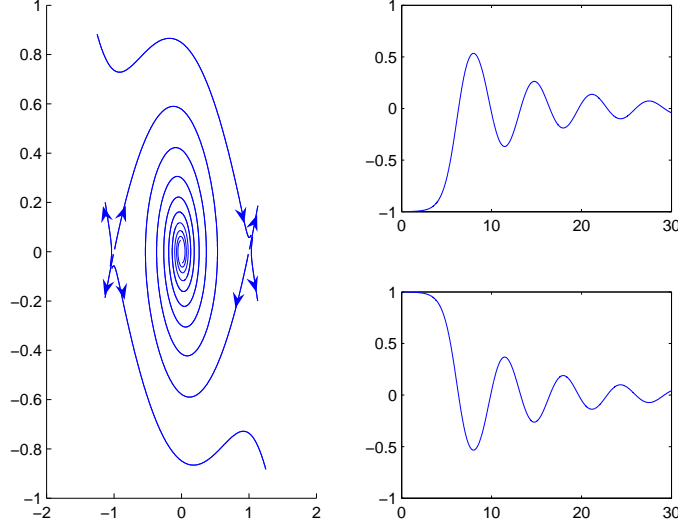


Figure 58: Phase plane for (3.6.21) with $c > 0$ (left): the traveling waves of interest must begin at either of the fixed points $(U, U') = (\pm 1, 0)$ and end at $(U, U') = (0, 0)$. Profiles of the heteroclinic orbits (right).

See [124, §§6.1-2] and [61, Chap. 6] for more information on propagating action potentials.

3.6.2 Traveling waves in the F-N equation

Rather than tackling the spatially-dependent H-H equations directly, we will consider the simplified F-N version. Adding the “diffusive” term $\frac{1}{pR} \frac{\partial^2 v}{\partial x^2}$ to (3.4.4a) and assuming constant resistance per unit length along the axon, we have:

$$\frac{\partial v}{\partial t} = \frac{1}{\tau_v} \left(\frac{1}{pR} \frac{\partial^2 v}{\partial x^2} + v - \frac{v^3}{3} - r + I_{app} \right), \quad (3.6.22a)$$

$$\frac{\partial r}{\partial t} = \frac{1}{\tau_r} (-r + 1.25v + 1.5). \quad (3.6.22b)$$

Recalling from §3.4 that $\tau_v \ll \tau_r$ we define a small parameter $\epsilon \stackrel{\text{def}}{=} \tau_v/\tau_r$. We then rescale time and space by letting $\bar{t} = t/\tau_r$ and $\bar{x} = (\epsilon\sqrt{pR})x$ so that $\frac{\partial}{\partial t} = \frac{1}{\tau_r} \frac{\partial}{\partial \bar{t}}$, $\frac{\partial}{\partial x} = (\epsilon\sqrt{pR}) \frac{\partial}{\partial \bar{x}}$ and (3.6.22) becomes

$$\epsilon \frac{\partial v}{\partial \bar{t}} = \epsilon^2 \frac{\partial^2 v}{\partial \bar{x}^2} + v - \frac{v^3}{3} - r + I_{app} \stackrel{\text{def}}{=} \epsilon^2 \frac{\partial^2 v}{\partial \bar{x}^2} + f(v, r), \quad (3.6.23a)$$

$$\frac{\partial r}{\partial \bar{t}} = (-r + 1.25v + 1.5) \stackrel{\text{def}}{=} g(v, r). \quad (3.6.23b)$$

This subterfuge involves choosing time and space scales in which the rise and fall of the action potential are equally steep in time and space (it also reduces the number of parameters by three). Since, as in §3.6.1, we want the length of the axon (cable) to be long in comparison to its diameter,

which is consistent with assuming a high resistance $R \gg 1$, the rescaling makes physical sense. (If the space domain in the original coordinates is $0 \leq x \leq L$, in the new ones it is $0 \leq \bar{x} \leq (\epsilon\sqrt{pR})L$.)

We seek a uniform traveling wave with voltage and recovery variable profiles $v(\bar{x}, \bar{t}) = V(\bar{x} - c\bar{t})$ and $r(\bar{x}, \bar{t}) = R(\bar{x} - c\bar{t})$. Proceeding as in Example 9, we obtain the pair of ODEs

$$\epsilon^2 V'' + \epsilon c V' + f(V, R) = 0, \quad (3.6.24a)$$

$$cR' + g(V, R) = 0, \quad (3.6.24b)$$

where $()' = \frac{\partial()}{\partial(\bar{x} - c\bar{t})}$. This is essentially a *third order* ODE whose analysis would take us too deeply into the technicalities of singular perturbation theory [15, 121], although it might make a nice project for the more mathematically-inclined). Briefly, since $\epsilon \ll 1$, the “waveframe dynamics” of the V -ODE is fast compared to that of R , so that we may assume R to remain approximately constant while V evolves rapidly towards a point on the nullcline $f(V, R) = 0$, according to (3.6.24a). After V equilibrates, R changes slowly according to (3.6.24b) and the solution moves along the nullcline until the bend, where it jumps rapidly to the other attracting branch, much as in Fig. 53. Keener and Sneyd [124, SS6.3-4] explicitly solve a version in which the cubic V -dependence in $f(V, R)$ is replaced by a piece-wise linear function representing the outer, negatively-sloped branches. They assemble exponential solutions by choosing constants of integration to satisfy matching conditions. They also sketch a geometrical approach to the nonlinear problem. Also see [61, §6.3].

3.7 Concluding remarks

In this section we have described relatively detailed *single compartment* models of isolated neurons that reproduce the dynamics of ionic channels. Many generalizations of the original Hodgkin-Huxley model of the squid’s giant axon have been developed, incorporating additional ionic currents. These typically involve multiple gating variables (sometimes ten or more) in addition to the membrane voltage, and analytical studies are not feasible. Numerical studies of such systems, with their multiple time scales, are also quite challenging. Reductions based on removal of fast variables are helpful, but even for the two dimensional cases described in §3.4 it is hard to find fixed points analytically.

A class of substantially simplified integrate-and-fire models has therefore been proposed, which neglect the details of spike dynamics, and only represent voltage variations during the refractory period. These are particularly useful in modeling networks of interconnected neurons, and we defer our discussion of them to the next section, after describing models for synaptic connections.

4 Synaptic connections and small networks (≈ 2.5 weeks)

We now move from single cell models to considering networks, starting with a review of models for chemical synapses, in which neurons communicate via release and reception of neurotransmitter molecules junctions, and of direct electrical connections or gap junctions.

4.1 Synapses and gap junctions: Towards networks

Synapses are the features in neurons that allow communication of signals with other neurons. There are more than 10^{11} neurons in the human brain, and on average there are 1000 synaptic connections from each neuron, and some neurons receive up to 10,000 inputs to the dendritic tree. Purkinje fibers in the cerebellum receive up to 100,000 inputs. Sherrington originally thought all synapses were electrical in nature, but in the 1920's, Loewi discovered the chemical synapse, showing that signals from the vagus nerve to the heart are conveyed by acetylcholine (ACh). There was a debate then as to whether synapses were all electrical or all chemical, and as the century continued, scientists realized that there are two distinct categories of synapses: chemical and electrical. Electrical synapses involve direct contact of cytoplasm of two distinct cells and allow simple depolarizing signals. Chemical synapses involve the release of neurotransmitter from one neuron and its reception at another. Chemical synapses are substantially slower than electrical synapses, but allow more complicated behavior. Synaptic plasticity is crucial to learning, since it allows connections among cells (and hence brain areas) to weaken or strengthen in response to experience. As synapses change strength, pathways change, as we hope yours will in this course.

4.1.1 Electrical Synapses

In electrical synapses, also called gap junctions, there is direct cytoplasm contact in gap junction channels, at which the neurons are only 3.5 nm apart. Gap junction channels are specialized protein structures, and each has a pore of approximate diameter 1.5 nm. Each channel is made up of a pair of hemichannels (or connexons), each of which is made of six connexins, which are proteins that span their respective cell membranes. The two hemichannels meet and create a single channel, which can open and close through the changing orientation of the connexins. The pore can allow dyes and tracking particles to pass from one neuron to another, allowing study of the networked neurons.

Electrical synapses provide fast, bidirectional communication, due to the electrotonic transmission. The change in postsynaptic potential depends directly on the size and shape of the change in presynaptic potential, as well as on the effective input resistance of the postsynaptic area. A small postsynaptic neuron or extension will have a high input resistance and respond more strongly to a small current through the channel. In general, electrical synapses are depolarizing (excitatory) through the transmission of the increased potentials corresponding to spikes, but if there is an action potential with a hyperpolarizing after-potential, that hyperpolarization will be seen postsynaptically as well. Bidirectionality results from the cytoplasmic continuity, although there are exceptions: certain channels can also close in response to voltage changes, permitting transmission of spikes in only one direction; these are called rectifying synapses. Also, to isolate damaged cells, gap-junction channels close in response to low cytoplasmic pH or high Ca^{2+} , both of which are signs of cell damage.

Electrical synapses appear where rapid response is required. Escape reflexes are a major example: in goldfish, the tail-flip giant neuron is connected by electrical synapse to sensory input, allowing a rapid burst of speed in response to threatening stimuli. Electrical synapses allow connection of large groups of neurons, so that multiple small cells can function as a larger unit. The resistance

of the coupled network is low, since the cells are all in parallel, so any input results in a small voltage change in the network, but once there is a sufficient input to cause a spike, all cells spike synchronously. Rapid electrical synapses thus allow group triggering in an “all-or-none” mode. For example, ink release in various marine snails is coordinated and synchronized by electrical synapses.

Modeling: Gap junctions are typically modeled as simple resistors, passing a current proportional to the voltage difference between the cells in contact. Hence for the current passing from the i th to the j th cell one adds to the internal ionic currents a term $I_{gap} = +\bar{g}_{gap}(v_i - v_j)$, where \bar{g}_{gap} represents the conductance of the junction and $(v_i - v_j)$ is the potential difference between cell i and cell j . For a pair of H-H type model cells, this leads to:

$$C_1 \dot{v}_1 = -I_{1,ion}(\dots) + I_{1,ext} + \bar{g}_{gap}(v_2 - v_1), \quad C_2 \dot{v}_2 = -I_{2,ion}(\dots) + I_{2,ext} + \bar{g}_{gap}(v_1 - v_2) \quad (4.1.1)$$

(plus equations for the gating variables of each cell). Note that the gap junction current term is conventionally shown as $+I_{gap}$. We consider a pair of simpler integrate-and-fire neurons coupled by both gap junctions and inhibitory synapses in §4.2.4.

4.1.2 Chemical Synapses

Much of the following discussion, including the notation in the modeling section, is drawn from [50, Chap. 5].

In chemical synapses, neurons are separated by synaptic clefts between presynaptic terminals or boutons (swellings on the axon) and postsynaptic dendritic spines. The 20-40 nm synaptic cleft is as wide or wider than the typical intracellular space. Following the arrival of an action potential, an influx of Ca^{2+} occurs through voltage-gated channels located in the active zone of the synapse. This causes the synaptic vesicles, in which neurotransmitter molecules are stored, to fuse with the cell membrane, thus opening and releasing their contents: a process called exocytosis. After the neurotransmitter molecules are released they diffuse across the synaptic cleft to reach postsynaptic receptors, which are proteins that span the postsynaptic cell membrane. Activation of these receptors opens or closes ion channels in the postsynaptic neuron, which, depending on the ion channel opened, can produce excitatory (depolarizing) or inhibitory (hyperpolarizing) effects.

This chain of processes obviously introduces communication delays, which range from 0.3 ms to several ms. Additional delays exist on the postsynaptic side, where there are two methods of opening channels. Ionotropic receptors have directly gated ion channels, allowing a fast response which lasts for several milliseconds. Ionotropic receptors typically feature ACh as the neurotransmitter and appear in neural circuits which require rapid behavior. Metabotropic receptors act as more indirect gates: receptors start a messenger cascade which eventually opens or closes the ion channels. These connections have slower responses, but their effects last seconds to minutes, depending on the messenger cascade properties. Norepinephrine acts in this way at synapses in the cerebral cortex.

Chemical synapses allow amplification of signals. One synaptic vesicle releases thousands of neurotransmitter molecules, which can open many ion channels and thereby depolarize a much larger cell than would be possible with gap-junction synapses. The result at the postsynaptic neuron does not depend on the neurotransmitter, but on the properties of the receptor circuit. ACh appears

in excitatory and inhibitory synapses. Glutamate and GABA (γ -aminobutyric acid) are major neurotransmitters in the brain. AMPA and NMDA are ionotropic excitatory receptor types for glutamate, with AMPA exhibiting significantly faster activation and deactivation. NMDA receptors additionally depend on the postsynaptic potential, which, as discussed later, has implications for synaptic plasticity and learning. GABA activates both ionotropic and metabotropic inhibitory receptors. These neurotransmitters act somewhat like hormones, but their effect is generally faster and much more precise. In the autonomic nervous system, however, specialized release and reception sites are not as prevalent, so the neurotransmitter acts more slowly and over a more diffuse area.

Modeling: As for gap junctions, neurotransmitter arrival at a synapse can be modeled as a current source in the postsynaptic neuron. The neurotransmitter binds to receptors, which results in ion channels opening and permitting a current. The major variable influencing the synaptic current then, is the conductance of the postsynaptic ion channels, which is typically modelled as the product of a maximal conductance, \bar{g}_s , with the probability of being open, P_s . P_s can represent the probability of a single channel being open, or the fraction of open channels in a larger group. Once the transmitter unbinds from the receptor, the channels close. Since \bar{g}_s is a constant, the dynamics occur in P_s . This can be modeled like the gating variables in the H-H equations (3.3.1):

$$\frac{dP_s}{dt} = \alpha_s(1 - P_s) - \beta_s P_s, \quad (4.1.2)$$

where α_s and β_s respectively determine the rates at which channels open and close (e.g. if the fraction of closed channels is $1 - P_s$, then the net opening rate is $\alpha_s(1 - P_s)$): see [52, p. 15] (cf. [50, p. 180]).

Channels typically open more rapidly than they close, so $\alpha_s \gg \beta_s$. The closing rate is generally assumed to be constant, but the opening rate α_s depends on the concentration of neurotransmitter in the synaptic cleft, and may be expressed as a constant $\bar{\alpha}_s$ multiplied by a factor describing how the transmitter concentration depends on the presynaptic voltage v_i , as in the expression

$$\alpha_s = \bar{\alpha}_s \frac{C_{NT,max}}{1 + \exp[-k_{pre}(v_i - E_{syn}^{pre})]} \stackrel{\text{def}}{=} \bar{\alpha}_s G(v_i), \quad (4.1.3)$$

from [52, p. 5]. In this sigmoidal function $C_{NT,max}$ represents the maximal neurotransmitter concentration in the synaptic cleft, k_{pre} sets the “sharpness” of the switch, and E_{syn}^{pre} sets the voltage at which it opens. However, due to the rapid rise of the AP, the transmitter fills the cleft to its full concentration faster than any other time constant in the model, so it behaves like an instantaneous switch that turns on and off as v_i passes up and down through E_{syn}^{pre} . We can therefore simplify by modeling the transmitter concentration as a step function that assumes a fixed value α_s for a fixed time T , after which it resets to 0:

$$\alpha_s(t) = \begin{cases} 0, & t \leq 0, \\ \alpha_s, & 0 < t < T, \\ 0, & t \geq T. \end{cases} \quad (4.1.4)$$

Under this approximation, if a presynaptic spike arrives at $t = 0$ when the postsynaptic open probability is $P_s(0)$, we can solve Equation (4.1.2) as a decoupled process, without further reference to the presynaptic cell’s state. Since $\alpha_s \gg \beta_s$, we may ignore β_s for $0 \leq t \leq T$, which gives

$$P_s(t) = 1 + (P_s(0) - 1) \exp(-\alpha_s t), \quad 0 \leq t \leq T. \quad (4.1.5)$$

If $P_s(0) = 0$, this rises to a maximum value $P_{max} = P_s(T) = 1 - \exp(-\alpha_s T)$ at time T . (This formulation also allows derivation of α_s experimentally, given the presynaptic voltage and the measured maximum conductance P_{max} .) For $t \geq T$, $\alpha_s = 0$, and we simply have exponential decay:

$$P_s(t) = P_s(T) \exp(-\beta_s(t - T)). \quad (4.1.6)$$

Putting (4.1.5) and (4.1.6) together, the postsynaptic current is approximated by an explicit piecewise-smooth function of time.

If the synaptic rise time is slower, such that the assumptions behind (4.1.5) do not hold, we can write the synaptic conductance after a presynaptic spike at $t = 0$ as the difference of two exponentials:

$$P_s(t) = P_{max} B(\exp(-t/\tau_1) - \exp(-t/\tau_2)), \quad t \geq 0. \quad (4.1.7)$$

The rise and fall in synaptic conductance is achieved by setting $\tau_1 > \tau_2$. B is a normalization factor to ensure that the maximum conductance is P_{max} .

Exercise 21. What is the correct value for B , in terms of τ_1 and τ_2 ?

Another expression with a similar form is the so-called “alpha function”

$$P_s(t) = \frac{P_{max} t}{\tau_s} \exp\left(1 - \frac{t}{\tau_s}\right), \quad t \geq 0, \quad (4.1.8)$$

which starts at zero, rises and peaks at the value of P_{max} at time $t = \tau_s$, and then decays back toward zero with a time constant τ_s . All the expressions (4.1.5-4.1.8) employ stereotypical functions $P_s(t)$ for the postsynaptic current that incorporate some physiological detail, but that are automatically elicited by a presynaptic spike, so that a train of such spikes occurring at times t_1, t_2, \dots produces

$$I_{syn}^{post} = \bar{g}_s \sum_i P_s(t - t_i). \quad (4.1.9)$$

If all the synaptic time constants are short then one may further simplify the analysis by replacing the functions $P_s(t)$ with delta functions⁹, possibly with a delay to account for neurotransmitter transport across the synaptic cleft, so that (4.1.9) becomes:

$$I_{syn}^{post} = \bar{g}_s \sum_i \delta(t - \tau_s - t_i); \quad (4.1.10)$$

we illustrate this in §4.2.4 below.

Other factors that arise in modeling synapses include postsynaptic voltage effects for NMDA receptors. If the postsynaptic neuron is near resting voltage, the channels will not open for a presynaptic spike. For the channels to open, there must be activity in both presynaptic and postsynaptic neurons. Once open, the NMDA receptors allow Ca^{2+} ions to enter, which increases the long-term synaptic strength. Such modification of synaptic responses to presynaptic spikes is called *synaptic plasticity*, and it includes both long-term effects as seen in NMDA receptors, and short-term effects, which are also known as synaptic facilitation and depression. *Synaptic facilitation* describes stronger postsynaptic response to a presynaptic spike that comes a short time after previous spikes,

⁹See §5.1.1 for a brief description of the delta function.

and *synaptic depression* is seen in a significantly lowered responses to each successive spike in a train. Short-term effects can be modeled by adding a second synaptic conductance term, P_{rel} , corresponding to the probability of neurotransmitter release at a presynaptic spike. In the absence of presynaptic spikes, P_{rel} decays to a resting value, as modeled by:

$$\tau_P \frac{dP_{rel}}{dt} = P_0 - P_{rel}. \quad (4.1.11)$$

For facilitation, after every presynaptic spike, we replace P_{rel} by $P_{rel} + f_F(1 - P_{rel})$, with $0 \leq f_F \leq 1$ representing the degree of facilitation. For depression, replace P_{rel} by $f_D P_{rel}$, also with $0 \leq f_D \leq 1$.

Finally, the synaptic current at the j th (postsynaptic) cell due to an AP in the i th (presynaptic) cell can be incorporated into the postsynaptic voltage equation as follows:

$$C\dot{v}_j = -I_{j,ion}(\dots) + I_{j,ext} - \bar{g}_{syn} P_{rel} P_s (v_j - E_{syn}^{post}). \quad (4.1.12)$$

In solving this equation we can either take $P_s = P_s(t - t_i)$ as a stereotypical function of the time elapsed since the i th cell spiked, as given by (4.1.5-4.1.6), (4.1.7) or (4.1.8). Alternatively, if we wish to include the dynamics of neurotransmitter release and represent dependence on presynaptic voltage more accurately, we can use the form (4.1.3) in (4.1.2) and solve Eqns. (4.1.12) and (4.1.2) as a coupled system, along with the gating equations. (This was done in, e.g., [76].)

For more discussion and a more detailed model of chemical synapses, see Keener and Sneyd [124, Chap. 7]. Every stage of synaptic transmission is covered there, from the presynaptic calcium currents to neurotransmitter diffusion. Exploring, analyzing, and simulating this model would be an interesting semester project. Johnston and Wu [120] also cover synapses in great detail (they occupy about one-third of the book). Chapter 11 concerns presynaptic mechanisms, Chapter 12 explores the role of calcium in transmitter release. Chapter 13 examines postsynaptic mechanisms, and Chapters 14 and 15 go into other synapse-related topics, with Chapter 15 investigating learning and memory.

4.2 Integrate-and-fire models of neurons and synapses

Much of the complexity of the H-H equations for a single-compartment neuronal model and reductions of them like the F-N model is due to the relatively detailed ionic channel (gating) dynamics responsible for the spike dynamics; namely, the variables m, n and h in Eqn. (3.3.1). These details are important in determining the shape and duration of the AP, but stereotypical APs are sufficient for many purposes, especially when they are effectively filtered by slower synaptic dynamics. Neglecting the details of spike generation allows us to significantly reduce the twenty-plus H-H parameters (cf. Eqns. 3.3.2-3.3.4), and to obtain a more tractable class of systems that can be more readily analyzed and simulated with less computer time and storage. These integrate-and-fire (IF) models allow us to probe the dynamics of larger networks, and to develop a range of models adapted to match different types of cortical neurons. In particular, we can fit specific input/output functions or current-to-frequency ($f - I$) curves more easily than by changing the gating dynamics, or adding further ionic currents. For example, initiation of spiking in typical cortical neurons is not via the Hopf bifurcation exhibited by the H-H model (but see the modified Connor-Stevens model [47, 48]).

IF models improve on the firing rate (connectionist) models that we encountered in §2, which are too crude to allow the study of spike trains. In this “middle-ground” approach we simplify the AP to a Dirac delta function, so that we need only model the refractory period during which the membrane voltage $v(t)$ builds toward threshold v_{th} . Upon reaching v_{th} at t_k , a delta spike $\delta(t - t_k)$ is inserted and $v(t)$ instantaneously reset at its minimum reference value v_r . The membrane voltage then begins to increase again, possibly after a brief absolute refractory period τ_{ref} . Unlike firing rate models, IF models generate spike trains and can thus reproduce a variety of spike-timing phenomena. Furthermore, as will be shown later, they allow inclusion of biophysical cellular and synaptic components of neural networks so that one may study their emergent behavior and effects on network function. IF models are therefore widely used in systems neuroscience. They can also be directly related to firing rate models, e.g. via approximation schemes such as the mean field approach [176].

Historically, IF models were proposed by Louis Lapicque in 1907 [137] who studied the sciatic nerve that excites leg muscles in frogs. Theoretical models were defined and analyzed decades later (e.g., see [204, 126, 127] and [30]).

4.2.1 Integrate-and-fire models for a single cell

The simplest IF model is the perfect integrator of [72]. The subthreshold membrane potential dynamics is described by the linear ODE

$$C\dot{v} = I_{syn} + I_{app}, \quad \text{for } v \in [v_r, v_{th}) \Rightarrow v(t) = v_r + \frac{1}{C} \int_0^t (I_{syn} + I_{app}) dt, \quad (4.2.1)$$

where I_{syn} is the total synaptic input from other neurons, and I_{app} is the intracellular injection current. A delta function of appropriate weight $w\delta(t)$ is superposed on v and the reset rule applied when $v(t)$ reaches v_{th} . The discontinuity (and effective nonlinearity) appears only when $v = v_{th}$. We discuss how to implement I_{syn} in §4.2.3. Fig. 59 shows the solution, when an absolute refractory period τ_{ref} is included and the input currents are constant in time and noise-free.

Exercise 22. Assuming that input currents are noiseless and constant in time, compute the interspike interval T_{ISI} (= spiking period) for the perfect IF neuron (4.2.1).

Lapicque’s [137, 30] work shows that the membrane potential is leaky, with a time constant $\tau = RC$ which is independent of stimulation area, and associates the leak with a parallel R-C circuit. In the 1960’s, Stein, Knight [204, 126] and others formally defined and analyzed such *leaky integrate-and-fire* (LIF) models in which the subthreshold dynamics is described by

$$C\dot{v} = -g_L(v - E_L) + I_{syn} + I_{app}, \quad \text{for } v \in [v_r, v_{th}), \quad (4.2.2)$$

The additional term in (4.2.2) includes the leak conductance, g_L , and resting potential, E_L , of the cell (E_L is the steady-state of v in the absence of input currents). Note that Eqn. (4.2.2) is precisely Eqn. (3.3.1a) less the ionic gating terms for spiking dynamics. Ignoring the reset rule, all solutions of (4.2.2) would approach an equilibrium $v_{ss} = E_L + (I_{syn} + I_{app})/g_L$. For $I_{syn} = 0$ and $I_{app} > g_L(v_{th} - E_L)$, v_{ss} lies above v_{th} and so the voltage will reach threshold and repetitive spiking will result. If $I_{app} < g_L(v_{th} - E_L)$ the voltage will settle at a subthreshold value $v_{ss} < v_{th}$,

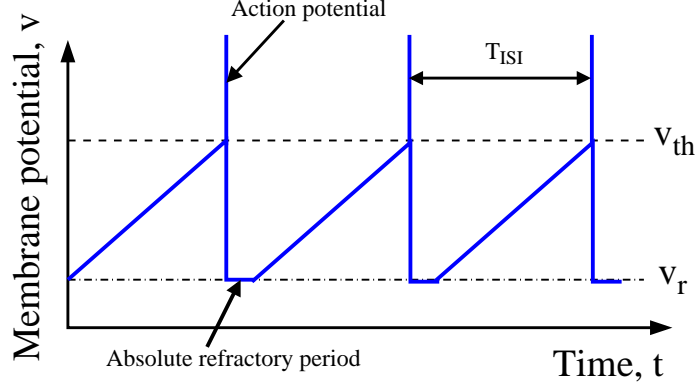


Figure 59: Dynamics of a noiseless perfect integrate-and-fire spiking model. When the membrane potential v reaches threshold v_{th} , a delta spike is inserted and immediately reset. A brief absolute refractory period is usually included. The firing rate is the inverse of the interspike interval $f = 1/T_{ISI}$.

and spikes cannot occur in the absence of noise. However, with the addition of sufficiently large noise, $v(t)$ can cross threshold with finite probability. Fig. 59 shows the solution of (4.2.2) in the noise-free case, with $v_{ss} > v_{th}$. In both of these examples analytical expressions for the interspike interval T_{ISI} (and hence for the firing rate $f = 1/T_{ISI}$), can be easily computed (see Exercises 22 and 23).

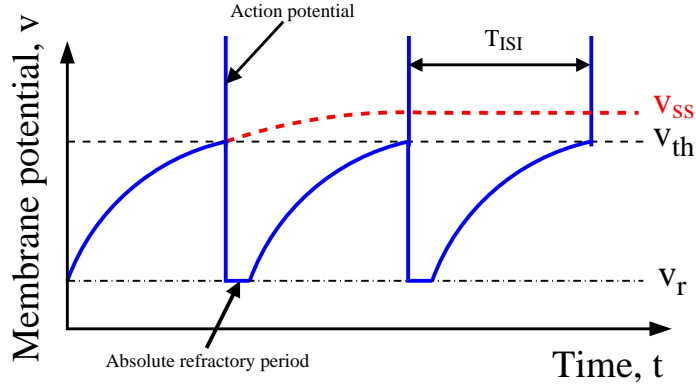


Figure 60: Dynamics of a noiseless leaky integrate-and-fire spiking model. The steady-state v_{ss} must exceed the threshold v_{th} for spiking to occur. Trajectory of $v(t)$ in the absence of the threshold shown in red.

Exercise 23. Show that, for $I_{syn} = 0$ and $I_{app} > g_L(v_{th} - E_L)$, T_{ISI} for the LIF model (4.2.2) is given by

$$T_{ISI} = \tau_{ref} + \frac{C}{g_L} \ln \left(\frac{v_{ss} - v_r}{v_{ss} - v_{th}} \right),$$

where the steady-state $v_{ss} = E_L + I_{app}/g_L$. How does the absolute refractory period τ_{ref} affect the shape of the $f - I$ curve? Plot f vs. I_{app} for fixed v_r, v_{th}, C, g_L, E_L and several τ_{ref} values to illustrate.

The LIF model has subthreshold dynamics that cause $v(t)$ to decelerate as it approaches v_{ss} , and hence as it approaches the threshold for spiking. This differs qualitatively from the accelerating

pre-spike dynamics of more realistic neuronal models such as the H-H equations (cf. Figs. 47 and 51a). The form of $v(t)$ can be improved by incorporating nonlinearity in the model, which becomes

$$C\dot{v} = F(v) + I_{syn} + I_{app}. \quad (4.2.3)$$

The simplest case is to make $F(v)$ a quadratic function: $F(v) \sim v^2$ [60, 138] or $F(v) = k_0(v - E_L)(v - v_c)$ [139], for $k_0 > 0$ and $v_{th} > v_c > E_L > v_r$. Suppose $I_{syn} = 0$. If $I_{app} = 0$ and $v < v_c$, $\dot{v} < 0$ and $v(t) \rightarrow E_L$, while if $v > v_c$, then $\dot{v} > 0$ and $v(t)$ will be automatically driven to threshold. Thus the unstable fixed point v_c acts as an effective threshold. When $I_{app} > 0$ exceeds a critical value, a saddle-node bifurcation occurs, $\dot{v} > 0$ for all $v \in [v_r, v_{th}]$ and periodic firing occurs. The QIF model is related to the θ -neuron model and the Type I neuron normal form [60, 56, 90].

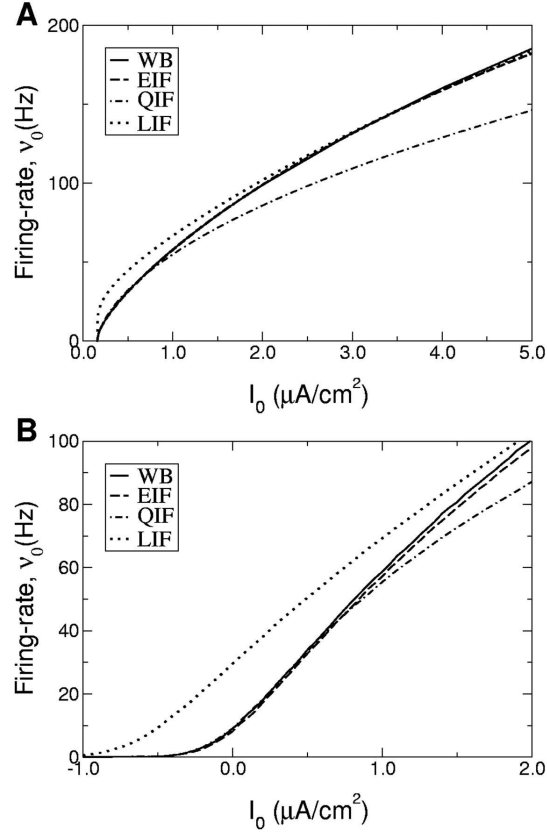


Figure 61: $f - I$ curves for different models are compared to each other and the more realistic Wang-Buzsaki (WB) model [227], which is a modified version of the H-H model. (A) Constant current inputs. (B) Noisy current inputs. Adapted from [66].

Yet more realistic waveforms can be constructed by using *exponential integrate-and-fire* (EIF) models [66, 65] where $F(v) \sim \exp((v - v_{th})/\Delta_{th})$ and Δ_{th} is a quantity that sets the sharpness of spike initiation. The top panel of Fig. 61 compares $f - I$ curves for various models.

Other IF models include the linear IF model [152, 69, 175] with a constant leak current, and the spike response model [73, 74] (that accounts for spiking latency) which is closer to a H-H model. Generalized IF models (GIF) [178, 29] are used to describe subthreshold resonance (see figure). See also the Abbott-Kepler model [2] reducing from H-H model to a nonlinear IF model. See [34, 35] for

model reviews and [115] for comparisons among them. See e.g. [94, 23, 22] for details on numerical methods for simulations of IF models.

4.2.2 Integrate-and-fire models with noisy inputs

Cortical neurons typically spike in an noisy and irregular manner. In reality, they (along with most other) neurons are continually bombarded by synaptic inputs in the form of small post-synaptic potentials (PSPs). Thus, I_{app} is not constant but more realistically takes the form $I_{app} + \eta(t)$, where I_{app} is the mean input level, which may be excitatory (positive) or inhibitory (negative), and $\eta(t)$ is a random process. Rewriting in terms of the noise-free steady state $v_{ss} = E_L + I_{app}/g_L$, Eqn. (4.2.2) becomes

$$\dot{v} = \frac{1}{\tau_m}(v_{ss} - v) + \frac{\eta(t)}{C}, \quad \text{for } v \in [v_r, v_{th}], \quad (4.2.4)$$

where $\tau_m = C/g_L$ is the time scale.

If we assume that the PSPs are much smaller than $|v_{th} - v_{reset}|$, $\eta(t)$ can be modeled as additive Gaussian noise and Eqn. (4.2.4) written as a *Langenvin equation*:

$$dv = (v_{ss} - v) \frac{dt}{\tau_m} + \frac{\sigma}{\sqrt{\tau_m}} dW(t), \quad (4.2.5)$$

where $dW(t)$ represents independent increments of a Wiener process or Brownian motion [71] (see Definition 9 in §6.1.2). We shall develop some of the probability and stochastic process theory necessary to better appreciate such processes in §§5-6.

Solutions of Eqn. (4.2.5) can be approximated by the Euler-Maruyama method [99], an extension of the forward Euler method of §2.3.1 to stochastic differential equations. Upon each timestep of length Δt the voltage variable is updated as follows:

$$v(t + \Delta t) = v(t) + \frac{\Delta t}{\tau_m}[v_{ss} - v(t)] + \sigma \sqrt{\frac{\Delta t}{\tau_m}} N(0, 1). \quad (4.2.6)$$

Here the notation $N(0, 1)$ indicates that independent samples are drawn from a Gaussian distribution with zero mean and unit variance. As in §2.3.1 the deterministic increment in Eqn. (4.2.6) is proportional to the nondimensional time step $\Delta t/\tau_m$, but the stochastic increment must be proportional to $\sqrt{\Delta t/\tau_m}$ to ensure that each increment of the Wiener diffusion process is normally distributed with mean zero and variance $\Delta t/\tau_m$. As $\Delta t \rightarrow 0$ successive iterates of (4.2.6) converge to a solution of Eqn. 4.2.5, although notions of convergence for random variables are more complicated than those discussed in §2.3, and must be formulated in terms of expectations. See [99] for an introduction with examples and Matlab scripts, and §6.4 below, for (a little) more background on stochastic differential equations. The Euler-Maruyama method is described in §6.4.4.

Exercise 24. Using Eqn. (4.2.6), investigate what happens when additive Gaussian noise is included in I_{app} ? How does the f - I curve change with increasing standard deviation σ ?

How do we quantify irregularity in spike rates and interspike intervals? The Fano factor, defined as the ratio $F = (\text{variance of ISI}) / (\text{mean of ISI})$ for a fixed time bin T , provides one measure of

variability (you will encounter this again in §5.1.2). Another quantity is the coefficient of variation, $C_v = (\text{standard deviation of ISI}) / (\text{mean of ISI})$. For a homogenous Poisson process, $F = 1$, independent of T . In a LIF model, we can create irregularity (large F) by driving a neuron by fluctuations, as in the next exercise.

Exercise 25. Use Eqn. (4.2.6) with fixed, large noise size σ to calculate and compare the C_v values for inputs of a spike train that follows a Gaussian distribution when the excitatory mean input current I_1 is (i) subthreshold ($v_{ss} < v_{th}$), (ii) superthreshold ($v_{ss} > v_{th}$). Also, (iii), consider combined excitatory ($2I_1$) and inhibitory ($-I_1$) inputs with net mean value I_1 but independent fluctuations both of standard deviation σ . Is the C_v in (iii) greater than that in (i)? Can you find values of σ and input currents (I_1) for which the C_v 's are greater than 1? (There is experimental evidence that networks can exhibit balances between excitation and inhibition; this is an ongoing area of research, see, e.g., [192, 211, 215, 5, 193, 28].)

The lower panel of Fig. 61 shows how the $f - I$ curves for different models change when noise is added, and Figs. 62 and 63 show examples of $f - I$ curves fitted to data from pyramidal and other cortical cells.

4.2.3 Implementation of synaptic inputs to IF models

There are two major approaches to implementing synaptic inputs: *current-based* and *conductance-based*. There are also two types of current inputs: excitatory and inhibitory, with reversal potentials E_e and E_i , respectively. For IF models we require $E_i < v_{reset} < v_{th} < E_e$.

Current-based IF models. In current-based IF models, the inputs are individual spikes generated by the presynaptic neuron(s), summed over the total number of presynaptic neurons or synaptic contacts, and over the times at which spikes occur. This simplest representation of this is to write the total input for the postsynaptic neuron i at time t as

$$I_{syn,i}(t) = \sum_j w_{i,j} \sum_k \delta(t - t_j^k), \quad (4.2.7)$$

where the index j labels the presynaptic neurons, $w_{i,j}$ is the synaptic weight from presynaptic neuron j to postsynaptic neuron i , and δ denotes the Dirac delta function which contributes a unit impulse ($\int_{-\infty}^{+\infty} \delta(t) dt = 1$) when neuron j spikes at times t_j^k . This implementation is a fairly good approximation if the synaptic channels open and close very rapidly. A temporal delay from presynaptic spikes due to signal transduction along axons or dendrites can be incorporated if necessary. The presynaptic spiking activity may include that from the local network or a stochastic background from other brain areas. Although this formulation simplifies mathematical analyses, it is a special case, as we shall see below.

A brief supplement on neurobiology of synapses Non-electronic receptors in synapses are categorized into ionotropic (ligand-gated ion channels) receptors or metabotropic receptors (which use second messengers). Ionotropic receptors can be further classified into three general classes: cys-loop receptors, ionotropic glutamate receptors, and ATP-gated channels. Here, we are interested only in the first two classes. Within the first class, it can be subdivided into the anionic and cationic groups. The most widely modeled for the former is Gamma-aminobutyric acid

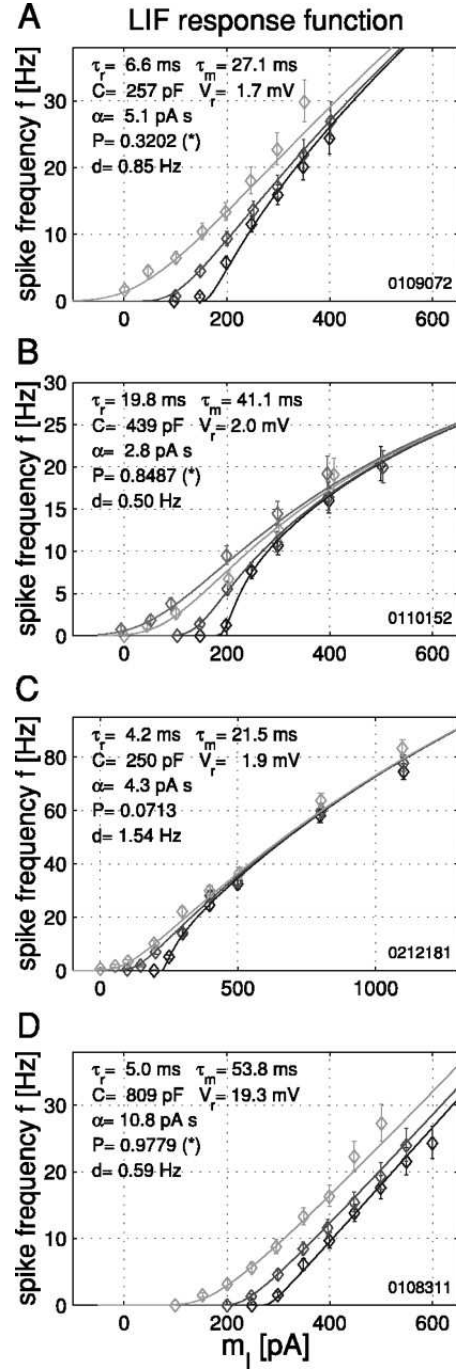


Figure 62: Fits of an LIF model to spike rate data from a pyramidal cell that receives in-vivo-like noisy input currents. Each panel uses different model parameters to fit data from four different cells. Adapted from [175].

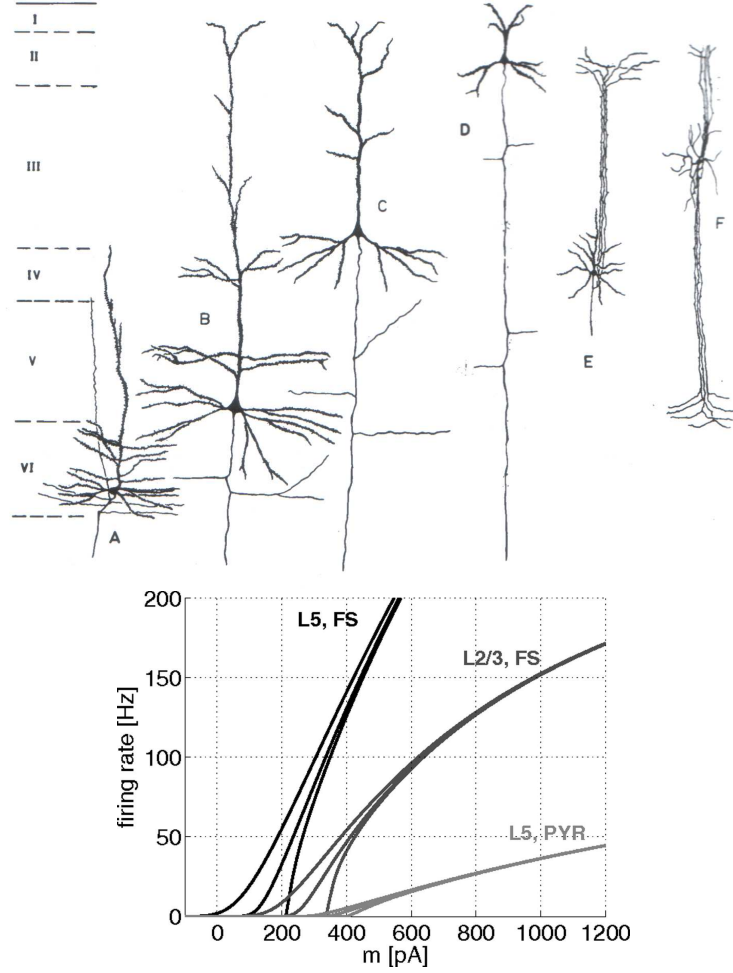


Figure 63: Neurons in different cortical layers I - VI can exhibit different morphologies and functional properties. Top: Cells labeled A to D are pyramidal cells, E is a spiny stellate cell and F a double bouquet cell. From [122]. Bottom: Diverse $f-I$ curves for cells in different layers; L2/3 and L5 refer to cortical layers (Roman numerals in top panel), FS denote fast-spiking inhibitory interneurons and PYR denotes excitatory pyramidal neurons. From [135].

(GABA_A) receptors, which is inhibitory. The latter are neuromodulatory receptors e.g. serotonin. The glutamatergic receptors consist of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA), N-methyl-D-aspartic acid (NMDA) and kainate receptors, which are excitatory to postsynaptic cells. Metabotropic receptors are usually more involved in its modeling, with a cascade of intracellular processes, and usually involving G-proteins. One of the more widely modeled G protein-coupled metabotropic receptor is gamma-aminobutyric acid (GABA_B), which is inhibitory. Here, we will not discuss about autoreceptors, which are receptors found on presynaptic sites. *Conductance-based* neuronal modeling involves the opening the closing dynamics of such receptors. The time constant of these receptors can be determined in *in vitro* experiments with the help of pharmacological blocking of specific receptors. For example, AP5, CNQX and bicuculline are so-called *antagonists* of NMDA, AMPA, and GABG_A receptors, respectively, by blocking their receptors. An alternative and powerful method is to “dynamically clamp” the conductances using a real-time interface between cell and a computer. See [169] for more discussions on such method.

Conductance-based IF models Here the synaptic conductance varies as a function of time, driven by the dynamics of auxiliary ODEs that are in turn driven by the presynaptic spike train. A single non-NMDA-mediated synapse (e.g. one using AMPA or GABA_A) produces an input current of the form

$$I_{syn,i}(t) = - \sum_j g_{i,j} S_j (v - E_j), \quad (4.2.8a)$$

$$\dot{S}_j = - \frac{S_j}{\tau_{S_j,decay}} + \alpha x_j(t), \quad (4.2.8b)$$

$$\dot{x}_j = - \frac{x_j}{\tau_{x_j,rise}} + \sum_k \delta(t - t_j^k), \quad (4.2.8c)$$

where the synaptic gating variable S is the fraction (probability) of open channels, $g_{i,j}$ is the peak synaptic conductance, E_j is the reversal potential for that particular ionic channel, and α is the weight contributed by each presynaptic spike. Alternatively, the synapse dynamics can be described by an alpha function.

Shunting inhibition When $E_i \approx E_L$ (e.g. with GABA_A), a phenomena called shunting inhibition can arise [110, 128], whereby the $f - I$ curve shifts laterally without change of shape when the inhibitory current increases. Because $v \sim gI$, early speculations proposed that increasing g by shunting inhibition can be thought of as changing the gain modulation of input to cells (i.e. the slope of the single-cell $f - I$ curves). However, Holt and Koch [110] showed that this is not the case. To see this, we add an inhibitory synaptic current to the LIF model:

$$\dot{v} = -g_L(v - E_L) - g_{syn}(v - E_i) + I_{app}, \quad (4.2.9)$$

and recompute the steady state

$$v_{ss} = \frac{g_L E_L + g_{syn} E_i + I_{app}}{g_L + g_{syn}}. \quad (4.2.10)$$

Defining the effective leak conductance $g_{L,eff} = g_L + g_{syn}$ ($> g_L$), we find that (cf. Exercise 23)

$$T_{ISI} = \tau_{ref} + \frac{C}{g_{L,eff}} \ln \left(\frac{v_{ss} - v_r}{v_{ss} - v_{th}} \right). \quad (4.2.11)$$

For large I_{app} $v_{ss} \gg v_{th}, v_r$ and we may use the approximation $\ln(1+x) \approx x$ to write

$$T_{ISI} \approx \tau_{ref} + \frac{C}{g_{L,eff}} \left(\frac{v_{th} - v_r}{v_{ss} - v_{th}} \right), \quad (4.2.12)$$

and if τ_{ref} is small enough to be ignored, we deduce that

$$f = 1/T_{ISI} \approx \left[\frac{g_{L,eff}(E_L - v_{th})}{C(v_{th} - v_r)} + \frac{I_{app}}{C(v_{th} - v_r)} \right]_+. \quad (4.2.13)$$

Clearly, increases in $g_{L,eff}$ only affect the first term on the right-hand-side of (4.2.13), which is independent of I_{app} . Hence, the $f - I$ curve shifts laterally without change of slope (= gain). Nonetheless, gain modulation can be achieved if synaptic noise [39] and/or dendritic properties [168, 38] are included.

When modeling AMPA- and GABA_A-mediated synapses with very short opening time constant $\tau_{x_j, rise}$, the auxiliary x variable is omitted, but this cannot be neglected for NMDA-mediated synapses. Not only are the time constants longer (especially in prefrontal cortex [222]), but the peak conductance also depends on the postsynaptic membrane potential v . (NMDA receptors are blocked by Mg^{++} ions, and the postsynaptic neuron has to be depolarized to activate this receptor [116, 117].) In this case the model takes the form

$$I_{NMDA,i}(t) = - \sum_j g_{i,j,NMDA} S_j (v - E_j), \quad (4.2.14a)$$

$$\dot{S}_j = - \frac{S_j}{\tau_{S_j, decay}} + (1 - S_j) \alpha x_j(t), \quad (4.2.14b)$$

$$\dot{x}_j = - \frac{x_j}{\tau_{x_j, rise}} + \sum_k \delta(t - t_j^k), \text{ where} \quad (4.2.14c)$$

$$g_{i,j,NMDA} = \frac{\bar{g}_{NMDA}}{\left[1 + \frac{[Mg^{++}]}{3.57} \exp(-v/16.13) \right]}. \quad (4.2.14d)$$

Note the nonlinearity (saturation) in S and voltage dependence in g_{NMDA} . The saturation can be appreciated if we assume regular presynaptic spiking ($\sum_k \delta(t - t_j^k) \sim$ firing rate $f \Rightarrow x_j \rightarrow \bar{x} = \tau_{x_j, rise}$). Then

$$\bar{S} = \frac{\tau f}{1 + \tau f}, \quad (4.2.15)$$

where $\tau = \alpha \tau_{x_j, rise} \tau_{S_j, decay}$.

Bistable single neuron With appropriate parameters, a model neuron with NMDA-mediated and additional currents and steady inputs can achieve an all-or-none activation pattern [142]. This is the simplest model of a bistable neuron. In Fig. 64 we plot the nonlinear current voltage relation given by (4.2.14a) and (4.2.14d) (with S_j constant) alone, and in combination with the linear relationship of (4.2.8a) for AMPA or GABA_A separately, and for all three currents together. The magenta curve for the combined currents intersects $I_{total} = 0$ three times, indicating the presence of two stable states separated by an unstable state for the ODE

$$C\dot{v} = I_{NMDA}(v) + I_{AMPA}(v) + I_{GABA_A}(v). \quad (4.2.16)$$

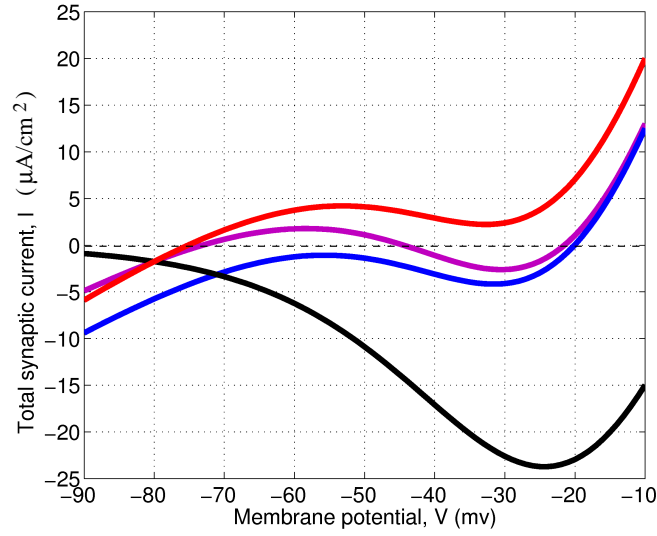


Figure 64: Illustration of bistable single-cell states for combined NMDA, GABA_A and AMPA currents via the current-potential or $I - v$ plot. NMDA synaptic currents alone cannot support bistability (black). When GABA_A currents are included, there is a single stable state (where I crosses zero; red). NMDA and AMPA currents also yield a single stable state (blue), but adding both GABA_A and AMPA currents can produce bistability (two stable states separated by an unstable state; magenta). Replotted using the values of [142, Fig. 1b].

Bistable neurons allow networks to robustly perform functions such as sensory integration [37, 142, 130, 81, 162]. Other models include nonlinear synaptic or dendritic dynamics (e.g. due to Ca^{++} , I_h -like, I_{ADP} and other currents) [143, 81, 161, 62, 145, 235]. The “up” and “down” states of bistable neurons and their spontaneous switching between these states have been observed experimentally in *in vivo* and *in vitro* neurons, see for example [205, 7, 199, 49, 145]. Other work with multiple stable states include [55, 67, 235].

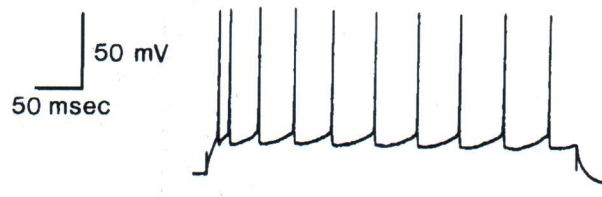


Figure 65: An example of spike frequency adaption in a pyramidal cell, showing membrane voltage under a constant applied current of 400 ms duration. Adapted from [151].

Other ionic currents and gap junctions There are a wide variety of ionic currents which we have not touched upon. They are often dependent on Ca^{2+} . A widely modeled synaptic currents include the widely-observed spike-rate or spike-frequency adaptation (afterhyperpolarization) phenomenon [153, 151] due to a calcium-dependent potassium current, which can be modeled as

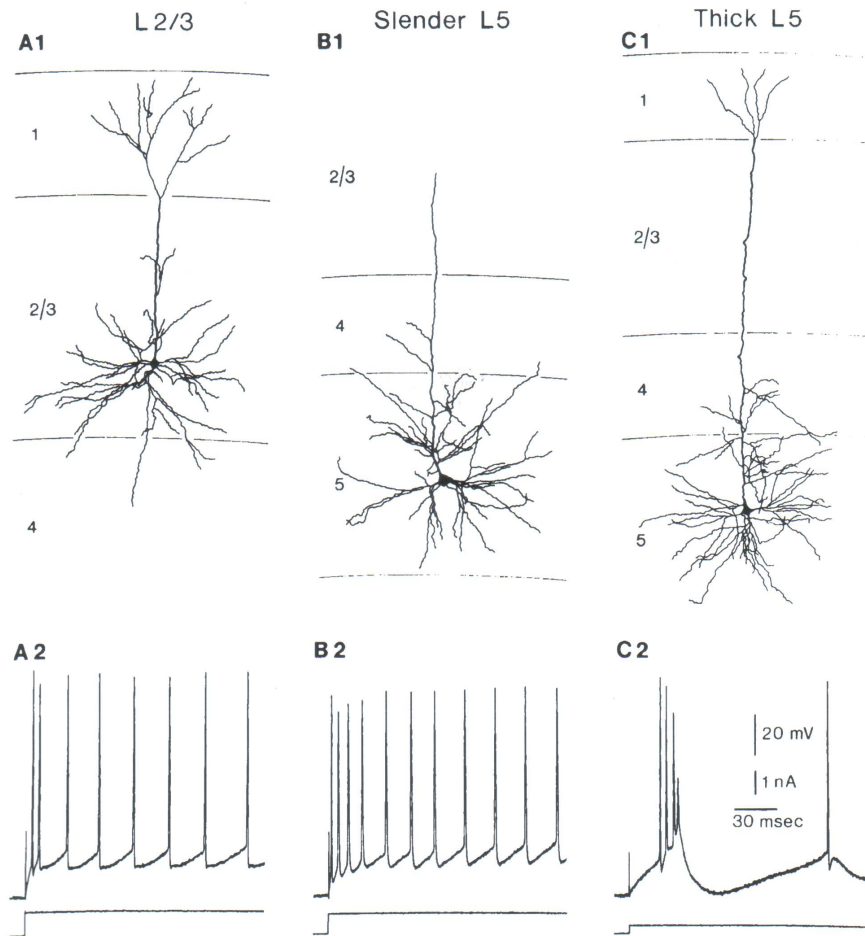


Figure 66: Diversity of spike frequency adaption in pyramidal cells, depending on their cortical layer. Adapted from [151].

[209, 210, 224, 144]

$$I_{syn,AHP}(t) = g_{AHP}[Ca^{2+}](v - E_K), \quad (4.2.17a)$$

$$[Ca^{++}] = \alpha_{Ca} \sum_k \delta(t - t_k) - \frac{[Ca^{++}]}{\tau_{Ca}}. \quad (4.2.17b)$$

As the calcium concentration $[Ca^{++}]$ increases due to an incoming spike train, the spike rate of the postsynaptic neuron falls: see Figs. 65-66 and [231, §10.1].

Gap junction or direct electrotonic coupling, described in §4.1.1 can also combine with chemical synapses to create bistable behaviors. See §4.2.4 for an example of two leaky integrate-and-fire neurons coupled via gap junctions.

In addition to postsynaptic conductance dynamics, presynaptic release of neurotransmitters can also be modeled. For example, it is known that presynaptic release of vesicles can exhibit short-term history dependence or plasticity that characterize changes in synapse strength. Short-term plasticity can be either facilitative or depressive [50].

Synaptic depression is due to depletion of neurotransmitter vesicles in the presynaptic cell. Letting $D(t)$ denote the fraction of available vesicles and appealing to the theory of kinetic processes [223], the dynamics of D can be modeled as:

$$\dot{D} = -\ln\left(\frac{1}{1-p_v}\right) \sum_k \delta(t - t_k) D + \frac{(1-D)}{\tau_D} \quad (4.2.18a)$$

$$\approx -p_v D \sum_k \delta(t - t_k) + \frac{(1-D)}{\tau_D}, \quad (4.2.18b)$$

where p_v is the release probability of each vesicle upon arrival of a spike, and τ_D (usually $\mathcal{O}(100)$ msec) sets the time scale for recharging vesicles after fusion and release. To incorporate this effect into the conductance models we simply multiply the factor α_x (amount of neurotransmitter release per spike) in Eqns. (4.2.8) and (4.2.14) by D . Synaptic facilitation, which is due to spike-triggered calcium processes, can be modeled with dynamics similar to those for depression [223]:

$$\dot{F} = \ln\left(\frac{1}{\alpha_F}\right) \sum_k \delta(t - t_k) (1 - F) + \frac{F}{\tau_F}, \quad (4.2.19a)$$

where $\tau_F \approx 100$ msec.

4.2.4 A pair of coupled integrate-and-fire neurons

Here we consider a pair of integrate-and-fire neurons coupled by both gap junctions and inhibitory synapses, with the effects of postsynaptic currents represented by delta functions. This results in simple linear ODEs that can be solved explicitly between spikes, allowing calculation of a Poincaré map. This example, which is analyzed in [141, 70], illustrates how a (fairly) simple one-dimensional map can be (almost) explicitly calculated, and how its structure reveals different dynamical regimes.

Writing the synaptic current in neuron i due to spiking in neuron j at $t = t_j^k$ as $I_{syn} = \bar{g}_{syn}(v_i - v_{syn}) \sum_k \delta(t - t_j^k)$ and using (4.1.1) in (4.2.2) results in the following system of two ODEs:

$$\begin{aligned} C_1 \dot{v}_1 &= -g_L(v_1 - v_L) + I_{1,ext} + \bar{g}_{gap}[(v_2 - v_1) + \Delta \sum_k \delta(t - t_2^k)] - \bar{g}_{syn}(v_1 - E_{syn}^{post}) \sum_k \delta(t - t_2^k), \\ C_2 \dot{v}_2 &= -g_L(v_2 - v_L) + I_{2,ext} + \bar{g}_{gap}[(v_1 - v_2) + \Delta \sum_k \delta(t - t_1^k)] - \bar{g}_{syn}(v_2 - E_{syn}^{post}) \sum_k \delta(t - t_1^k), \end{aligned} \quad (4.2.20)$$

where the delta function terms account for current flow due to potential differences across the gap junction during the (instantaneous) spikes, as well as the postsynaptic current. In (4.2.20) the $(v_i - v_j)$ components are called *subthreshold* and the delta function components, *superthreshold*. The parameter Δ quantifies the “weight” that the spike contributes to the voltage differences across the gap junction.

For simplicity, suppose that the cells are identical ($C_1 = C_2 = C$) and subject to the same external current $I_{1,ext} = I_{2,ext} = I_{ext}$. In that case we can rescale time by $g_L t / C \mapsto t$ and voltage by $(v - v_0) / (v_{th} - v_0) \mapsto v$ to map the range between reset v_0 and threshold v_{th} onto the unit interval $[0, 1)$. This gives:

$$\begin{aligned} \frac{dv_1}{dt} &= I - v_1 + \alpha [(v_2 - v_1) + \Delta \sum_k \delta(t - t_2^k)] - \beta(v_1 - v_{syn}) \sum_k \delta(t - t_2^k), \\ \frac{dv_2}{dt} &= I - v_2 + \alpha [(v_1 - v_2) + \Delta \sum_k \delta(t - t_1^k)] - \beta(v_2 - v_{syn}) \sum_k \delta(t - t_1^k). \end{aligned} \quad (4.2.21)$$

where α , β , I and v_{syn} are rescaled analogs of \bar{g}_{gap} , \bar{g}_{syn} , I_{ext} and E_{syn}^{post} .

Exercise 26. *Work out the details of the rescaling transformations noted above and find expressions for α , β , I and v_{syn} in terms of the original parameters.*

In the absence of spiking (or if $\Delta = \beta = 0$), and with initial conditions $v_1(0) = \bar{v}_1$, $v_2(0) = \bar{v}_2$, Eqns. (4.2.21) may be solved to give

$$\begin{aligned} v_1(t) &= I(1 - e^{-t}) + \left(\frac{\bar{v}_1 + \bar{v}_2}{2}\right) e^{-t} + \left(\frac{\bar{v}_1 - \bar{v}_2}{2}\right) e^{-at}, \\ v_2(t) &= I(1 - e^{-t}) + \left(\frac{\bar{v}_1 + \bar{v}_2}{2}\right) e^{-t} - \left(\frac{\bar{v}_1 - \bar{v}_2}{2}\right) e^{-at}, \end{aligned} \quad (4.2.22)$$

where $a = 1 + 2\alpha$. Hence $v_1 - v_2 = (\bar{v}_1 - \bar{v}_2)e^{-at}$: under the influence of subthreshold coupling (or indeed just leakage alone), the membrane voltages approach one another. However, spikes and resets to zero intervene whenever v_1 or v_2 reach the rescaled threshold value 1, so that we have a *hybrid dynamical system* rather than the smooth ODEs that we have encountered so far. To construct long time solutions of (4.2.21) we must assemble segments of solutions punctuated by jumps at appropriate points. This suggests that we derive a discrete mapping somewhat different from the one introduced in §2.4.4.

Assuming that cell 2 has just fired and the system is at $(v_1, v_2) = (\bar{v}_1, 0)$, we can compute a Poincaré map from this post-spike state to the next post-spike state immediately after cell 1 fires: indeed, from the first of Eqns. (4.2.22), the interspike interval (ISI) $T = T(\bar{v}_1)$ is determined by:

$$v_1(T) = I(1 - e^{-T}) + \frac{\bar{v}_1}{2} (e^{-T} + e^{-aT}) = 1, \quad (4.2.23)$$

and from the second equation the state of cell 2 when cell 1 fires is

$$v_2(T) = I(1 - e^{-T}) + \frac{\bar{v}_1}{2} (e^{-T} - e^{-aT}) = 1 - \bar{v}_1 e^{-aT}. \quad (4.2.24)$$

Superthreshold electrical coupling adds a further amount $\alpha\Delta$ to $v_2(T)$, and synaptic coupling subtracts $\beta(v_2(T) - v_{syn})$. Similar remarks apply in the case that cell 1 fires followed by cell 2. Hence, letting v denote the voltage of the cell that has *not* just fired immediately after the other one has fired, we obtain the spike-to-spike Poincaré map and its derivative:

$$v \mapsto f(v) = (1 - \beta)(1 - ve^{-aT}) + d(\alpha, \beta, \Delta, v_{syn}), \quad (4.2.25)$$

$$\frac{\partial f(v)}{\partial v} = f'(v) = -(1 - \beta)e^{-aT}(1 - avT'), \quad (4.2.26)$$

where $T = T(v)$ is determined by (4.2.23) and we have collected the synaptic and electrical “spike” terms in the parameter $d = (1 - \beta)\alpha\Delta + \beta v_{syn}$ which superimposes a vertical post-spike jump that shifts f upward if superthreshold electrical coupling dominates synaptic inhibition and downward otherwise. Fig. 67 shows three qualitatively different forms that the Poincaré map can exhibit as parameter values change.

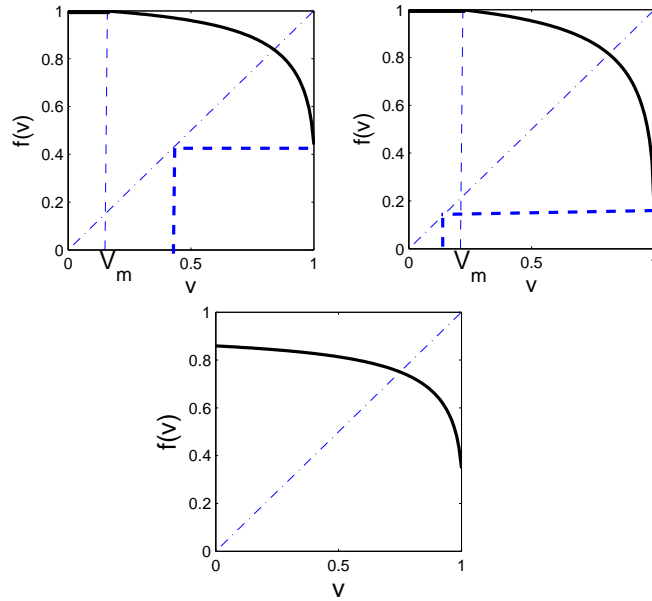


Figure 67: Three distinct forms for the Poincaré map $f(v)$ with coexisting synchronous and asynchronous states, showing disjoint domains of attraction ($0 < v_m < f(1^-)$: top left); interleaved domains of attraction ($0 < f(1^-) < v_m$: top right), and an isolated synchronous state ($v_m = 0$: bottom). Parameter values are (a) $I = 1.15$, $\beta = 0$, $\Delta = 0.05$; (b) $I = 1.19$, $\beta = 0$, $\Delta = 0.1$; (c) $I = 1.15$, $\beta = 0.1$, $\Delta = 0.05$, $v_{syn} = -0.5$ and $\alpha = 0.2$ throughout.

In interpreting these results $v = 1$ and $v = 0$ should be identified, since on spiking v resets instantaneously to 0. Moreover, since the spike form is compressed into a delta function, we must decide how to deal with the cases that a positive impulse carries v above threshold ($v = 1$) or a negative one takes it below reset ($v = 0$). In the former we assume that a spike occurs immediately: thus, if cell 1 fires when cell 2 is at $v_2(T)$ and $(1 - \beta)v_2(T) + d \geq 1$, cell 2 fires immediately and thereafter 1 and 2 fire in perfect synchrony: this is called *spike-capture synchrony* (cf. [141]), and it appears in the Poincaré map as a flat (horizontal) segment over the subinterval $[0, v_m]$. In the latter case we allow v_2 to drop below 0, so that v_1 fires successive spikes and v_2 's spikes are *suppressed*.

In all three cases shown in Fig. 67 there are coexisting stable fixed points at $v = 0 (= 1)$ and at v^e in the interior of the interval $(0, 1)$, but the nature of the separatrix dividing their domains of

attraction differs. Depending upon the size and sign of the post-spike jump d in (4.2.25), the values of v_m and the limit $f(1^-)$ of the map may take any of the three qualitatively distinct forms. In the first two there is a (small) domain $[0, v_m]$ in which spike-capture synchrony occurs. In the first case the interval $[0, v_m]$ is not accessible from the invariant domain $(v_m, 1)$ and all solutions starting in $(v_m, 1)$ limit either on the asynchronous fixed point v^e or on an orbit of period 2. In the second case, some solutions starting in $(v_m, 1)$ can escape to $[0, v_m]$, resulting in spike capture. These two cases are both shown without synaptic coupling ($\beta = 0$), but they can also occur for sufficiently small β and $\alpha > 0$. The third case occurs only with inhibitory synapses present ($\beta > 0$): here the synchronous state $v = 0$ is completely isolated and a positively-invariant domain surrounds v^e , and if $(1 - \beta)(1 + \alpha\Delta) + \beta v_{\text{syn}} < 0$ (synaptic coupling is sufficiently strong relative to electrical coupling), spike suppression occurs.

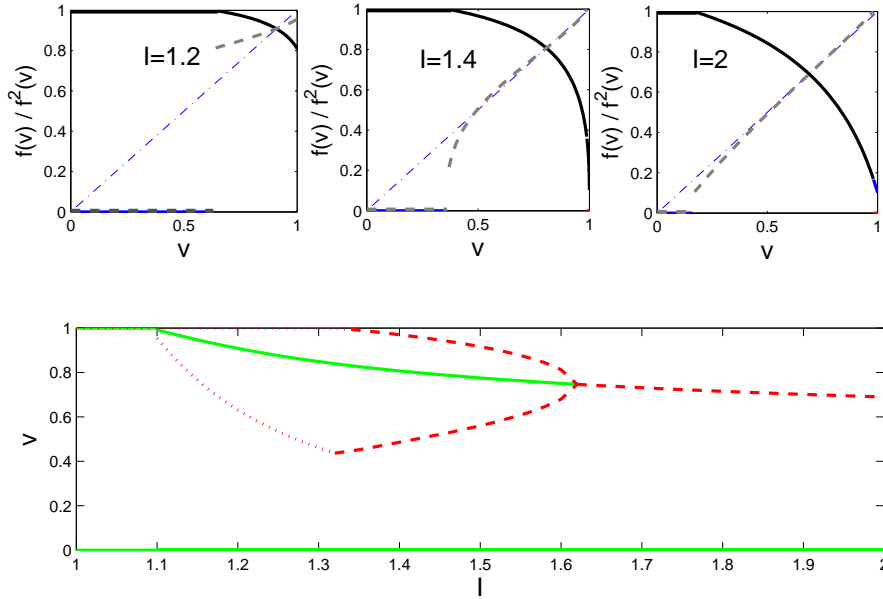


Figure 68: Poincaré maps and a bifurcation diagram for a pair of cells with gap junction coupling and delta function synapses. The top panels show $f(v)$ (bold) and $f^2(v)$ (dashed) for bias currents $I = 1.2, 1.4$ and 2 . The bottom panel shows branches of asynchronous fixed points v^e and unstable period 2 orbits arising in a subcritical period-doubling bifurcation and disappearing at the discontinuity in f when $v_m = f(1^-)$. A stable synchronous state ($v = 0 = 1$) exists throughout and is globally stable for $1 \leq I \leq 1.1$. Parameter values: $\alpha = 0.4$, $\Delta = 0.4$, $\beta = 0.05$, $v_{\text{syn}} = -1$.

As the applied current I increases the asynchronous fixed point v^e can lose stability in a period-doubling bifurcation [100, 89] in which unstable period 2 orbits shrink down on the branch of fixed points (a subcritical bifurcation). Laborious perturbation calculations for weak coupling (small α, β) confirm the stability properties of v^e illustrated in the bifurcation diagram of Fig. 68. The period 2 orbits and, for low bias current I , the discontinuity at $f(1^-)$, separate the domains of attraction of the synchronous and asynchronous states.

In §6.2 we describe a large network model for decision making, composed of 1400 excitatory

and 400 inhibitory integrate-and-fire cells, modeling a microcircuit within a cortical column, and describe how it can be reduced to a leaky accumulator model similar to that of Example 7.

4.2.5 A final note on neuronal modeling

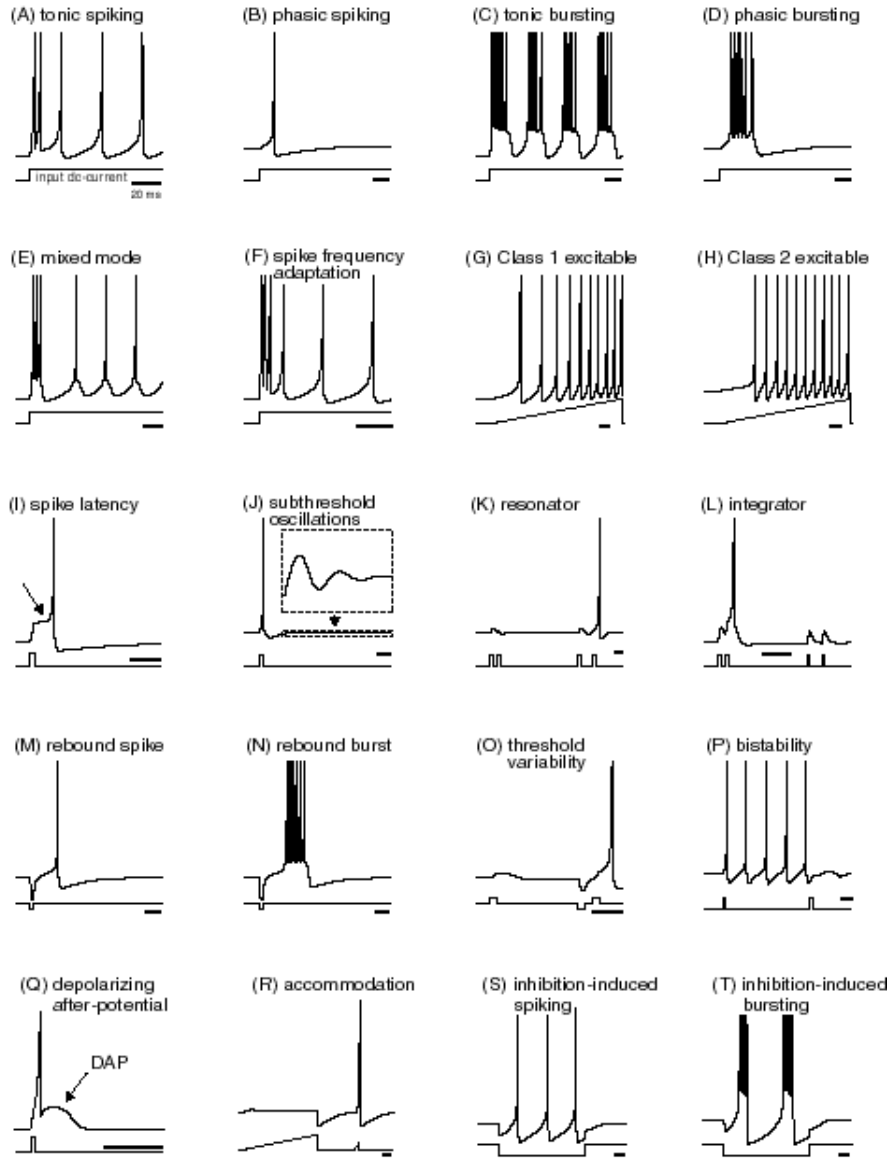


Figure 69: Diversity of spiking behavior in neurons. From [115].

We end by noting that parameters such as C , τ_m , g_{syn} , g_L , E_L , etc. can be measured or estimated from *in vitro* experimental studies, and are thus not “free” parameters. However, these and other parameters can vary widely depending on cell type and locations, as illustrated above in Figs. 62, 63 and 66. Fig. 69 summarises the rich variety of spiking and bursting patterns that have been observed. Thus, care must be taken in choosing parameters, especially when comparing with experimental data.

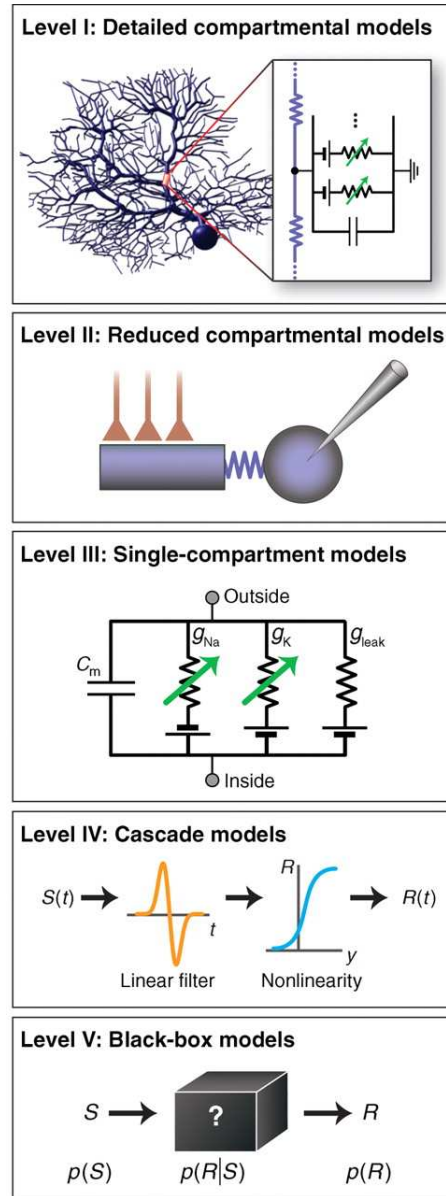


Figure 70: General classes of neuronal models for single neurons, with level of detail and complexity decreasing from top to bottom. From [98].

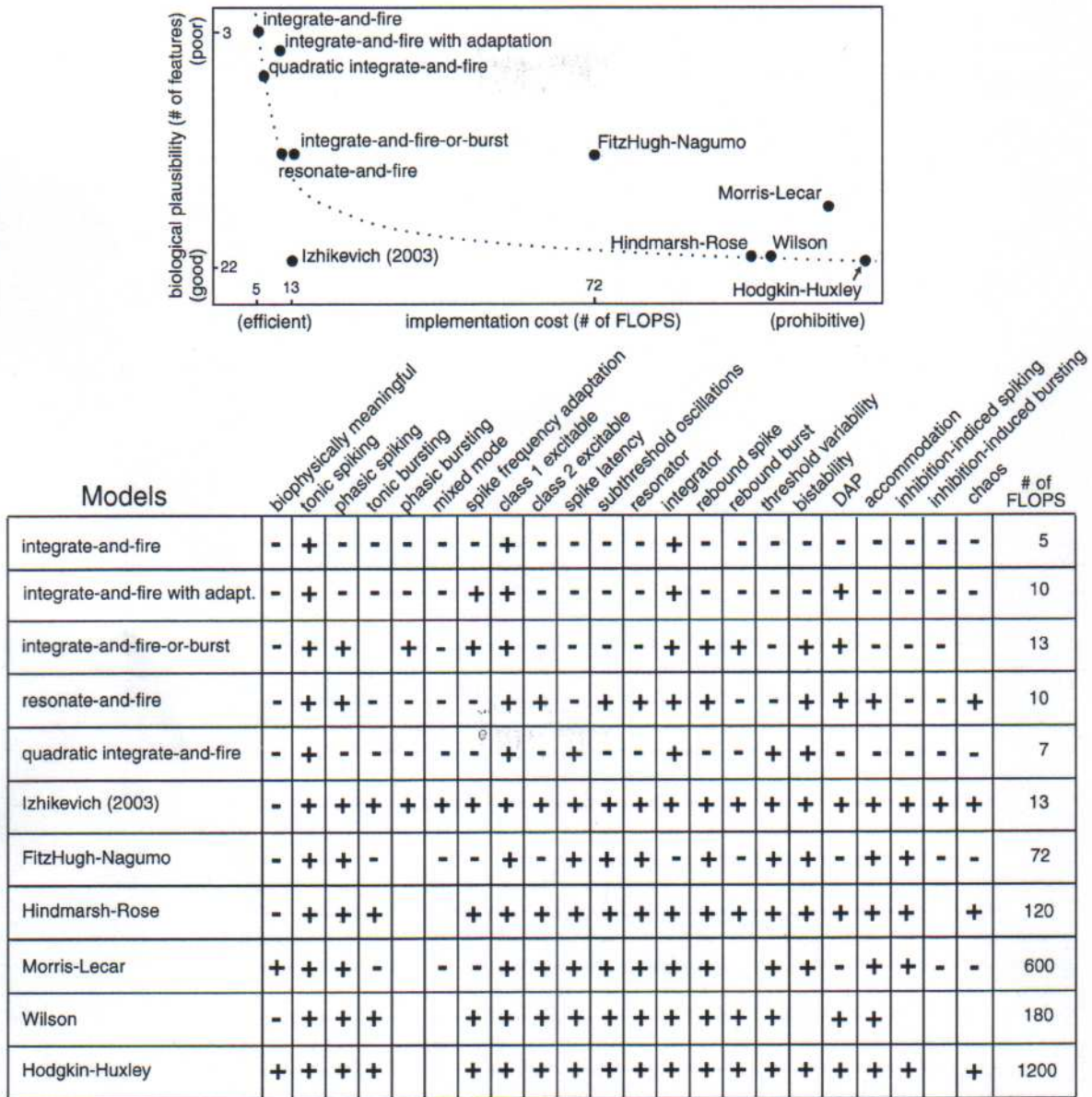


Figure 71: Comparison of computational costs for simulating different single compartment models. Features preserved or neglected by each model are indicated, and the last column shows the approximate number of floating point operations required to simulate the model for 1 ms. Some models (e.g. integrate-and-fire-or-burst, resonate-and-fire) are not included. From [115].

In §3 and the present section we have developed a range of models, from the relative complexity of H-H and related ionic current descriptions, with detailed spike dynamics, to IF models with their stereotypical delta function spikes. Fig. 70 summarises the range of spatial scales and complexities. In practice, synaptic currents and model types should be selected from the range of available neuronal models based upon the types of questions to be addressed [98]. Analytical tractability, and the ability to extract understanding of how parameters influence behavior through bifurcations, is one consideration. A second is computational tractability (Fig 71) [115].

4.3 Phase reductions and phase oscillator models of neurons

We have already noted that the presence of multiple ionic currents, each necessitating one or more gating variables, makes models of the form (3.3.1) analytically intractable, but that time scale separation allows one to reduce the dimension of state space, sometimes to two dimensions, so that phase plane methods can be used. In addition to reductions based on equilibrating fast gating variables and eliminating correlated ones (§3.4), and the integrate-and-fire simplifications of §4.2 that replace the spike dynamics by delta functions and stereotypical postsynaptic potentials, there is a third class of reduced models, based on the notion of phase. In these *phase oscillator* or *rotator* models, a single “clock-face” variable that tracks each cell’s progress toward spiking replaces the membrane voltage and gating variables.

4.3.1 Phase response curves

The ability to reduce multidimensional dynamics to a single phase variable is based on state space topology. Spontaneously spiking or bursting neuron models typically possess hyperbolic (exponentially) attracting limit cycles [100, 89]. In Euclidean space \mathbb{R}^n ODEs must have at least two state variables ($n \geq 2$) to possess limit cycles (§2.4.4), and so it might seem that the two-dimensional reductions of §3.4 are the best we can do. However, circular and toroidal phase spaces – manifolds – naturally allow orbits to return near their starting points. A limit cycle is, topologically, circular: it can be deformed into a geometrical circle by a nonlinear change of coordinates, no matter what ambient dimension it inhabits. Phase reduction uses this fact. The method was originally developed by Malkin [147, 148], and independently, with biological applications in mind, by Winfree [234]; also see [59, 56, 113].

Near a hyperbolic limit cycle, Γ_0 , of period T_0 , the n -dimensional state space locally splits into a phase variable ϕ along the closed curve Γ_0 , and a foliation of transverse *isochrons*: $(n - 1)$ -dimensional manifolds M_ϕ with the property that any two solutions starting on the same leaf M_{ϕ_0} are mapped by the flow to another leaf M_{ϕ_1} and hence approach Γ_0 with the same asymptotic phase [88] (see Figs. 72 and 73). To get (more) explicit expressions, we first write the H-H type (voltage + ionic gating) equations (3.5.2) in the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \epsilon \mathbf{g}(\mathbf{x}, \dots); \quad \mathbf{x} \in \mathbb{R}^n, \quad (4.3.1)$$

where $\mathbf{g}(\mathbf{x}, \dots)$ represents external (e.g., synaptic) inputs. Let $\phi \in [0, 2\pi)$ denote the phase coordinate around Γ_0 and the vector-valued function $\mathbf{x}_0(\phi)$ describe the location of Γ_0 in the state space for $\epsilon = 0$. We usually fix the “origin” so that the spike peaks at $\phi = 0$ and choose ϕ such

that $\dot{\phi} = \omega_0 = 2\pi/T_0$: progress around Γ_0 occurs at constant speed with respect to ϕ . Hence equal increments in ϕ correspond to equal time increments, and they are usually spaced non-uniformly around Γ_0 , as in the clustered isochrons of Fig. 73 below.

The fact that the isochrons form a smooth foliation implies that there is a smooth function $\phi(\mathbf{x})$ assigning phase values to points in a neighborhood of Γ_0 . Employing the chain rule, we may therefore obtain an ODE that describes the dynamics on the perturbed limit cycle Γ_ϵ :

$$\dot{\phi} = \omega_0 + \epsilon \frac{\partial \phi}{\partial \mathbf{x}} \cdot \mathbf{g}(\mathbf{x}_0(\phi), \dots) |_{\Gamma_0(\phi)} + \mathcal{O}(\epsilon^2). \quad (4.3.2)$$

Here we implicitly assume that coupling and external influences are weak ($\epsilon \ll 1$), so that Γ_0 perturbs to a nearby hyperbolic limit cycle Γ_ϵ , allowing us to compute the scalar phase equation by evaluating $\mathbf{g}(\mathbf{x}, \dots)$ along Γ_0 .

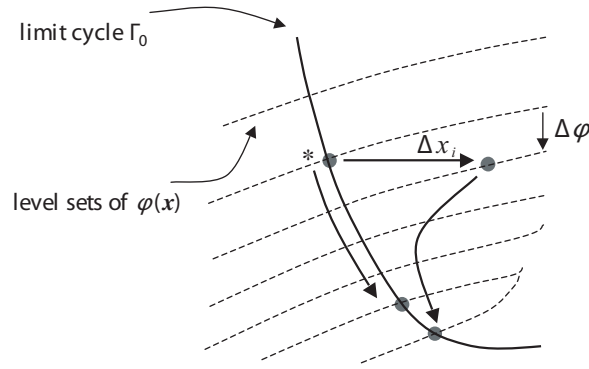


Figure 72: The direct method for computing the PRC, showing the geometry of isochrons, the effect of the perturbation that results in a jump to a new isochron, recovery to the limit cycle, and the resulting phase shift. Adapted from [26].

For neural models in which inputs and coupling enter only via the voltage equation (3.5.2a), $\frac{\partial \phi}{\partial v} \stackrel{\text{def}}{=} z(\phi)$ is the only nonzero component in the vector $\frac{\partial \phi}{\partial \mathbf{x}}$. This function $z(\phi)$ is called the *phase response curve* (PRC), and it describes the sensitivity of the system to inputs as a function of phase on the cycle. More biologically, it describes how impulsive (spike-like) perturbations advance or retard the next spike as a function of the time (\sim phase) during the cycle at which they are applied.

The PRC may be calculated directly by appeal to the usual finite-difference approximation to the derivative:

$$z(\phi) = \frac{\partial \phi}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{x}}{\partial V} = \frac{\partial \phi}{\partial V} = \lim_{\Delta V \rightarrow 0} \left[\frac{\phi(\mathbf{x}_0 + (\Delta V, 0)^T) - \phi(\mathbf{x}_0)}{\Delta V} \right], \quad (4.3.3)$$

where the numerator $\Delta\phi = [\phi(\mathbf{x}_0 + (\Delta V, 0)^T) - \phi(\mathbf{x}_0)]$ describes the change in phase due to a finite impulsive (delta function) perturbation $V \rightarrow V + \Delta V$ from a base point $\mathbf{x}_0 \in \Gamma_0$; see Fig. 72. (Since it is defined in the limit $\Delta V \rightarrow 0$, some authors call $z(\phi)$ the *infinitesimal PRC* (iPRC) [159].) Numerical simulations of solutions of the unperturbed ($\epsilon = 0$) ODE (4.3.1) starting at a set of well-chosen points \mathbf{x}_0 and perturbed from them by (small) ΔV yield acceptable results. Asymptotic approximations of PRCs may also be obtained near local and global bifurcations at which periodic spiking begins, using explicit solutions of ODEs written in normal form coordinates: we give an example at the end of this subsection. A recent collection of articles appears in [191].

There is also an adjoint formulation from which $z(\phi)$ may be computed numerically, as is done in the software package XPP [57]. This is also based on linearization around the limit cycle Γ_0 . Letting $\mathbf{x}_0(t)$ represent a solution of the unperturbed ODE on Γ_0 and $\Delta\mathbf{x}(t)$ an *arbitrary* small perturbation to it resulting from an impulsive perturbation such as ΔV , substitution of $\mathbf{x} = \mathbf{x}_0 + \Delta\mathbf{x}$ into the ODE yields

$$\frac{d}{dt}\Delta\mathbf{x}(t) = \mathbf{Df}(\mathbf{x}_0(t))\Delta\mathbf{x}(t) + \mathcal{O}(\|\Delta\mathbf{x}\|^2) : \quad (4.3.4)$$

a linear system of ODEs but with a time-dependent (periodic) Jacobian matrix (cf. linearization at a fixed point: Eqn. (2.1.1)). As in (4.3.3) but allowing the more general perturbation $\Delta\mathbf{x}$, the resulting phase shift may be approximated as

$$\Delta\phi = \frac{\partial\phi}{\partial\mathbf{x}} \cdot \Delta\mathbf{x} + \mathcal{O}(\|\Delta\mathbf{x}\|^2), \quad (4.3.5)$$

where $\mathbf{u} \cdot \mathbf{v}$ denotes the inner product in \mathbb{R}^n . The effects of the instantaneous perturbation $\Delta\mathbf{x}$ depend only on the place $\mathbf{x}_0 \in \Gamma_0$ at which it is applied, so that $\Delta\phi$ is independent of time. Hence we may take time derivatives in (4.3.5) to yield

$$\frac{d}{dt} \left(\frac{\partial\phi}{\partial\mathbf{x}} \right) \cdot \Delta\mathbf{x} \approx -\frac{\partial\phi}{\partial\mathbf{x}} \cdot \frac{d}{dt}\Delta\mathbf{x} \approx -\frac{\partial\phi}{\partial\mathbf{x}} \cdot \mathbf{Df}(\mathbf{x}_0(t))\Delta\mathbf{x} = -\mathbf{Df}^T(\mathbf{x}_0(t)) \frac{\partial\phi}{\partial\mathbf{x}} \cdot \Delta\mathbf{x}, \quad (4.3.6)$$

where we use (4.3.4) and in the final step appeal to the definition of the adjoint operator (the transposed matrix) in switching the order in which \mathbf{Df} appears. Since $\Delta\mathbf{x}$ is arbitrary, and ignoring the small $\mathcal{O}(\|\Delta\mathbf{x}\|^2)$ terms, (4.3.6) implies that

$$\frac{d}{dt} \left(\frac{\partial\phi}{\partial\mathbf{x}} \right) = -\mathbf{Df}^T(\mathbf{x}_0(t)) \frac{\partial\phi}{\partial\mathbf{x}}. \quad (4.3.7)$$

We have obtained a linear ODE whose solution is the generalized, vector-valued PRC! One initial condition for this ODE is provided by the observation that, for $\epsilon = 0$,

$$\dot{\phi} = \frac{\partial\phi}{\partial\mathbf{x}} \cdot \frac{d\mathbf{x}}{dt} = \frac{\partial\phi}{\partial\mathbf{x}} \cdot \mathbf{f}(\mathbf{x}) = \omega_0, \quad (4.3.8)$$

the remaining $n - 1$ come from the requirement that the solution $\frac{\partial\phi}{\partial\mathbf{x}}$ of (4.3.7) be T_0 -periodic. See [113, §9.1] for more information.

After the PRC is computed the phase model takes the form

$$\dot{\phi} = \omega_0 + \epsilon z(\phi)h(\phi, \dots) + \mathcal{O}(\epsilon^2), \quad (4.3.9)$$

where (\dots) in the argument of the input or coupling function h depends on how the cell is externally or synaptically perturbed. Since the cell itself is now described only through a single phase variable ϕ , and its phase $\phi = \omega_0 t$ moves around the circle at constant speed in the uncoupled limit $\epsilon = 0$, this ODE is sometimes called a *rotator*.

Example 10. *Fig. 73 shows an example of isochrons and PRCs computed for a two-dimensional reduction due to Rose and Hindmarsh [182] of a multi-channel neuron model of Connor et al. [46]:*

$$\begin{aligned} C\dot{v} &= [I^b - g_{Na}m_\infty(v)^3(-3(w - Bb_\infty(v)) + 0.85)(v - E_{Na}) \\ &\quad - g_K w(v - E_K) - g_L(v - E_L) + I^{ext}] \\ \dot{w} &= (w_\infty(v) - w)/\tau_w(v), \end{aligned} \quad (4.3.10)$$

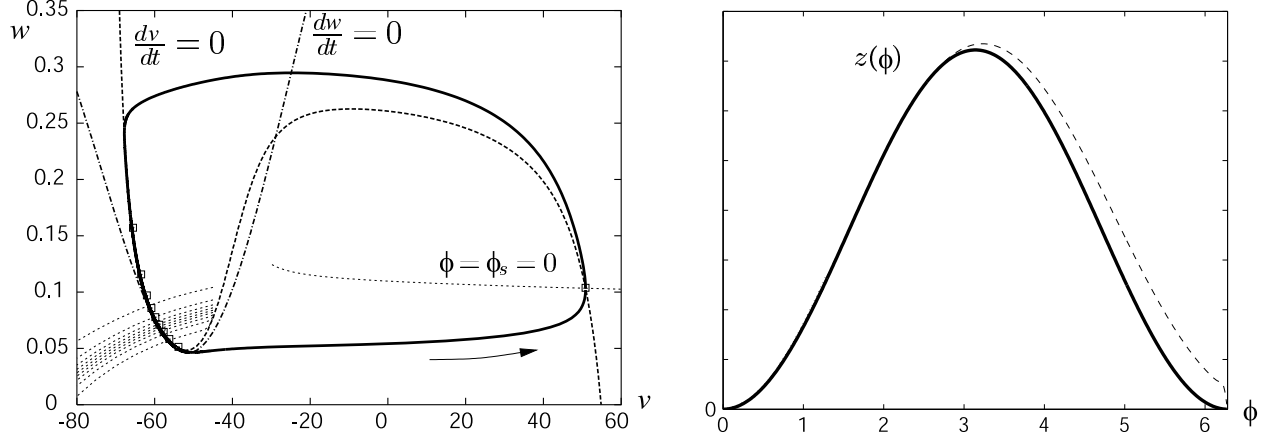


Figure 73: (a) The state space structure for a repetitively spiking Rose-Hindmarsh model, showing attracting limit cycle and isochrons. The dashed and dash-dotted lines are nullclines for $\dot{v} = 0$ and $\dot{w} = 0$, respectively, and squares show points on the perturbed limit cycle, equally spaced in time, under a small constant input current I_{ext} . (b) PRCs for the Rose-Hindmarsh model; the asymptotic form $z(\phi) \sim [1 - \cos \phi]$ is shown solid, and numerical computations near the saddle node bifurcation on the limit cycle yield the dashed result. From [27].

where the functions $m_\infty(v)$, $b_\infty(v)$, $w_\infty(v)$ and $\tau_w(v)$ are of the sigmoidal forms of (3.3.1-3.3.5). Since the gating variables have been reduced to a single scalar w by use of the timescale separation methods of §3.4, the isochrons are one-dimensional arcs. Note that these arcs, equally-spaced in time, are bunched in the refractory region in which the nullclines almost coincide and flow is very slow. In fact as the bias current I^b is reduced the $\dot{v} = 0$ nullcline moves downward and touches the $\dot{w} = 0$ nullcline, creating a saddle-node bifurcation on the closed orbit of (4.3.10). Thereafter the nullclines intersect transversely at a sink and a saddle. The sink corresponds to a quiescent, hyperpolarized state, and the saddle and its stable manifold provide the threshold for spiking.

The right hand panel of Fig. 73 shows the results of a direct numerical computation using (4.3.3) (dashed curve) and the analytical approximation derived from the normal form for a saddle-node bifurcation: $z(\phi) \sim [1 - \cos \phi]$ (solid curve).

The explicit approximation for the PRC in this example is derived from the normal form of the saddle-node bifurcation (recall Eqn. (2.4.13) in §2.4.3), which we write in a slightly different form to indicate that we are interested in the situation just *after* the fixed points have disappeared, i.e.

$$\dot{x} = \mu + x^2 \quad (4.3.11)$$

for small $\mu > 0$. This separable ODE may be integrated to give $x(t) = \sqrt{\mu} \tan(\sqrt{\mu}(t - t_0))$: a solution that blows up to ∞ as the elapsed time $t - t_0 \rightarrow \pi/2\sqrt{\mu}$ and comes from $-\infty$ as $t - t_0 \rightarrow -\pi/2\sqrt{\mu}$. Identifying the singularity with the “time” at which the cell fires¹⁰, this gives a period $T_0 = \pi/\sqrt{\mu}$ and frequency $\omega_0 = 2\sqrt{\mu}$. The time derivative is therefore

$$\dot{x} = \frac{\mu}{\cos^2(\sqrt{\mu}(t - t_0))}. \quad (4.3.12)$$

¹⁰A tricky move! We’ll try to justify this by drawing pictures in class.

Adopting the convention that the spike occurs at $\phi = \omega t = 0$, we have $\sqrt{\mu}(0 - t_0) = -\pi/2 \Rightarrow t_0 = \pi/2\sqrt{\mu}$, so that the solution and its time derivative may be rewritten as:

$$\dot{x} = \frac{\mu}{\cos^2(\sqrt{\mu}t - \pi/2)} = \frac{\mu}{\sin^2(\sqrt{\mu}t)} = \frac{\omega_0^2}{2[1 - \cos(2\sqrt{\mu}t)]} = \frac{\omega_0^2}{2[1 - \cos \phi]}. \quad (4.3.13)$$

Here we replace the parameter μ by ω_0 using $\omega_0 = 2\sqrt{\mu}$ and also use the fact that the uncoupled rotator (4.3.9) is solved by $\phi = \omega_0 t$. Finally we compute, as above but with only scalar functions:

$$z(\phi) = \frac{d\phi}{dx} = \frac{d\phi}{dt} \frac{dt}{dx} = \frac{\omega_0}{\dot{x}} = \frac{2}{\omega_0}(1 - \cos \phi). \quad (4.3.14)$$

In this saddle-node-in-a-periodic-orbit case (sometimes called SNIPER: an awful acronym!), positive perturbations can only advance the phase and hasten the next spike. In other cases, phase can be advanced or retarded by positive perturbations; for example, the PRC for the small limit cycle near a Hopf bifurcation is approximately sinusoidal [58, 26].

4.3.2 Phase models and averaging

Suppose that we have computed the phase reduction for a repetitively spiking neuron and wish to study how two such identical cells interact when synaptically or electrically coupled. We now have a pair of equations like (4.3.9):

$$\dot{\phi}_1 = \omega_0 + \epsilon[\delta_1 + z(\phi_1)h_1(\phi_1, \phi_2)] \stackrel{\text{def}}{=} \omega_0 + \epsilon H_1(\phi_1, \phi_2), \quad (4.3.15a)$$

$$\dot{\phi}_2 = \omega_0 + \epsilon[\delta_2 + z(\phi_2)h_2(\phi_2, \phi_1)] \stackrel{\text{def}}{=} \omega_0 + \epsilon H_2(\phi_2, \phi_1); \quad (4.3.15b)$$

in writing this we allow small frequency differences (detuning) $\epsilon\delta_j$ and we neglect the yet smaller $\mathcal{O}(\epsilon^2)$ terms. The state space of this system is the product of two circles: a two-dimensional torus around which solutions wind with approximately the same speed in each direction ($\phi_1 = \omega_0 t + \mathcal{O}(\epsilon)$). To proceed further we change variables to remove this common frequency, defining new phase variables $\psi_i = \phi_i - \omega_0 t$, so that (4.3.15) becomes:

$$\dot{\psi}_1 = \epsilon H_1(\psi_1 + \omega_0 t, \psi_2 + \omega_0 t), \quad (4.3.16a)$$

$$\dot{\psi}_2 = \epsilon H_2(\psi_2 + \omega_0 t, \psi_1 + \omega_0 t). \quad (4.3.16b)$$

Equations (4.3.16) are now in a form to which we can apply the averaging theorem [89, §§4.1-2], which may be somewhat informally stated as follows:

Theorem 4. *Consider the following ODEs*

$$\dot{\mathbf{x}} = \epsilon \mathbf{f}(\mathbf{x}, t), \quad \text{where } \mathbf{f}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t + T) \text{ is } T\text{-periodic in } t, \text{ and} \quad (4.3.17)$$

$$\dot{\mathbf{y}} = \epsilon \overline{\mathbf{f}(\mathbf{y})}, \quad \text{where } \overline{\mathbf{f}(\mathbf{y})} = \frac{1}{T} \int_0^T \mathbf{f}(\mathbf{y}, t) dt \text{ is the average of } \mathbf{f}(\mathbf{x}, t). \quad (4.3.18)$$

Then solutions $x(t)$ of (4.3.17) and $y(t)$ of (4.3.18), if started within $\mathcal{O}(\epsilon)$, remain within $\mathcal{O}(\epsilon)$ for times of $\mathcal{O}(1/\epsilon)$. Moreover, hyperbolic fixed points of (4.3.18) correspond to hyperbolic T -periodic orbits of (4.3.17) with the same stability types.

In computing the integral of (4.3.18) one treats the state variable vector \mathbf{y} as constant, and only averages the explicitly t -dependent terms in the vectorfield $\mathbf{f}(\mathbf{y}, t)$. For more details and a proof, see [89, §4.1].

Recalling that the common period of the unperturbed rotators is T_0 , we compute the averages of the terms on the RHS of (4.3.16):

$$\overline{H_1(\psi_1, \psi_2)} = \frac{1}{T_0} \int_0^{T_0} H_j(\psi_1 + \omega_0 t, \psi_2 + \omega_0 t) dt, \quad (4.3.19)$$

and similarly for $\overline{H_2(\psi_2, \psi_1)}$.

In interpreting the resulting averaged vectorfield, we use the fact that the H_j 's are necessarily 2π -periodic in each of the phase variables. Specifically, changing variables in the integral of (4.3.19) by letting $\tau = \psi_2 + \omega_0 t$, so that $dt = \frac{d\tau}{\omega_0} = \frac{T_0 d\tau}{2\pi}$, it becomes

$$\overline{H_1(\psi_1, \psi_2)} = \frac{1}{2\pi} \int_0^{2\pi} H_1(\psi_1 - \psi_2 + \tau, \tau) d\tau \stackrel{\text{def}}{=} \overline{H_1(\psi_1 - \psi_2)}, \quad (4.3.20)$$

and we find that the averaged functions $\overline{H_1(\psi_1 - \psi_2)}$ and $\overline{H_2(\psi_2 - \psi_1)}$ depend only upon the *difference* between the slow phases! These functions are also necessarily 2π -periodic in the phase difference.

4.3.3 Phase models and half-center oscillators

For mutually-symmetric coupling between two identical cells we have $h_2(\phi_1, \phi_2) = h_1(\phi_2, \phi_1)$ in the original functions of (4.3.15). Equations (4.3.15) are symmetric under permutation of ϕ_1 and ϕ_2 : what cell 1 does to cell 2, cell 2 does to cell 1. This implies that $\overline{H_2(\psi_1 - \psi_2)} = \overline{H_1(\psi_2 - \psi_1)} \stackrel{\text{def}}{=} \overline{H(\psi_2 - \psi_1)}$ after averaging, and that the reduced phase equations are also permutation symmetric:

$$\dot{\psi}_1 = \epsilon \overline{H(\psi_1 - \psi_2)} \quad \text{and} \quad \dot{\psi}_2 = \epsilon \overline{H(\psi_2 - \psi_1)}. \quad (4.3.21)$$

We may subtract these to further reduce to a single scalar ODE for the phase difference $\theta = \psi_1 - \psi_2$:

$$\dot{\theta} = \epsilon [\overline{H(\theta)} - \overline{H(-\theta)}] \stackrel{\text{def}}{=} \epsilon G(\theta). \quad (4.3.22)$$

Since the function H is 2π -periodic, we have $G(\pi) = \overline{H(\pi)} - \overline{H(-\pi)} = \overline{H(\pi)} - \overline{H(\pi)} = 0$ and $G(0) = \overline{H(0)} - \overline{H(0)} = 0$, implying that, *regardless of the precise form of \overline{H}* , in-phase and anti-phase solutions *always* exist. Moreover, because $G(-\theta) = [\overline{H(-\theta)} - \overline{H(\theta)}] = -G(\theta)$, G is an odd function and its derivative $G'(\theta)$ is even: see Fig. 75 below. Additional fixed points can also exist, depending on the details of \overline{H} . In general, if there is a phase difference θ^e such that $G(\theta^e) = 0$, we say that $\theta = \theta^e$ is a *phase-locked* solution.

The stability of phase-locked solutions is determined by the eigenvalues of the 2×2 matrix obtained by linearizing (4.3.21) at $\psi_1 - \psi_2 = \theta^e$:

$$\epsilon \begin{bmatrix} \overline{H(\theta^e)}' & -\overline{H(\theta^e)}' \\ -\overline{H(-\theta^e)}' & \overline{H(-\theta^e)}' \end{bmatrix}; \quad (4.3.23)$$

these are 0 and $\epsilon[\overline{H(\theta^e)}' + \overline{H(-\theta^e)}'] = \epsilon G'(\theta^e)$, with eigenvectors $(1, 1)^T$ and $(\overline{H(\theta^e)}', -\overline{H(-\theta^e)}')^T$ respectively. (For $\theta^e = 0$ and π the second of these is $(1, -1)^T$.) Hence the dynamics is only neutrally stable to perturbations that advance or retard the phases of both units equally, but if $\overline{H(0)}' < 0$ (resp. $\overline{H(\pi)}' < 0$) then the in-phase (resp. anti-phase) solution is asymptotically stable to perturbations that disrupt the relative phase $\psi_1 - \psi_2$.

Note that θ is also the difference between the original phase variables: $\theta = \phi_1 - \phi_2$. Thus, if (4.3.22) has a stable fixed point at $\theta = \theta^e$, then there is a periodic solution of the original system in which the phases approximately maintain the difference θ^e , implying that spikes alternate with a regular spacing. If $\theta^e = 0$, we say that spikes are *synchronized*; if $\theta^e = \pi$, they are *antisynchronized*. Terms such as frequency locking are also used to describe solutions in which the phase difference $\theta = \phi_1 - \phi_2$ remains constant. See [231, §12.1] for additional discussions.

Together, phase reduction and averaging have simplified a system that contains at least two voltage variables, two gating variables, and possibly two synaptic variables, to a one-dimensional system on the phase circle.

Exercise 27. Consider the following (averaged) symmetrically-coupled system:

$$\dot{\psi}_1 = \omega_1 - \alpha \sin(\psi_1 - \psi_2), \quad \dot{\psi}_2 = \omega_2 - \alpha \sin(\psi_2 - \psi_1).$$

Find conditions on the parameters ω_j and α that imply frequency locking, investigate the stability of frequency locked solutions and express their phase difference(s) $\phi_1 - \phi_2$ in terms of the frequency difference $\omega_1 - \omega_2 \stackrel{\text{def}}{=} \Delta\omega$ and coupling strength α . What happens when frequency locking is lost?

Exercise 28. Modify the system of Exercise 27 to read:

$$\dot{\psi}_1 = \omega_1 - [\sin(\psi_1 - \psi_2) - \beta \sin(2(\psi_1 - \psi_2))], \quad \dot{\psi}_2 = \omega_2 - [\sin(\psi_2 - \psi_1) - \beta \sin(2(\psi_2 - \psi_1))].$$

Compute the reduced system of the form (4.3.22) and analyze the bifurcations that occur as β increases from zero for fixed $\Delta\omega = \omega_1 - \omega_2$ sufficiently small that phase locked solutions exist near 0 and π .

A common structure appearing in (idealized models of) central pattern generators is the *half-center oscillator*: a reflection-symmetric pair of units, sometimes each containing several neurons, that are usually connected via mutual inhibition to produce an alternating rhythm. The reduction to two phase oscillators described above is perhaps the simplest expression of this architectural subunit. It will reappear in the next section.

4.4 Central Pattern Generators

Central pattern generators (CPGs) are networks of neurons in the spinal cords of vertebrates and invertebrate thoracic ganglia, capable of generating muscular activity in the absence of sensory feedback (e.g. [43, 75, 164, 149], cf [231, Chaps. 12-13]). Studies of locomotion generation are perhaps most common, but CPGs drive many other rhythmic activities, including scratching, whisking (e.g. in rats), moulting (in insects), chewing and digestion (indeed, the stomato-gastric ganglion in lobster is perhaps the best-modeled among them [150]). CPGs are typically studied in preparations

isolated *in vitro*, with sensory inputs and higher brain “commands” removed [43, 84], and sometimes in neonatal animals, but it is increasingly acknowledged that an integrative approach, including muscles, body-limb dynamics and proprioceptive feedback is needed to fully understand their function [40, 107, 212]. CPGs nonetheless provide examples of neural networks capable of generating interesting behaviors, but small enough to allow the study of relatively detailed biophysically-based models. Here we sketch two examples.

4.4.1 A CPG model for lamprey

As §4.1 shows, the direct study of a pair of synaptically (or electrically) coupled H-H type neurons, or even of two-dimensional reductions thereof, is challenging, so much of the CPG modeling has been done at a higher, more phenomenological level. Phase models have been particularly useful. A relatively early CPG model for lamprey swimming [41] used a chain of N phase oscillators with nearest-neighbor coupling to describe the distributed networks of cells in the spinal cord that generate traveling waves (also see [58, 129]). At that time it could not be derived from single cell models or otherwise justified, but following the discovery of distinct cell types [33] and the creation of cell-based [85, 96, 220] and network-based models [32, 230], its status has been strengthened by computing PRCs and using phase reduction, as in §4.3 [216].

The phase oscillator model of [41] takes the form:

$$\begin{aligned}\dot{\theta}_1 &= \omega_1 + \alpha \sin(\theta_2 - \theta_1), \\ \dot{\theta}_i &= \omega_i + \alpha \sin(\theta_{i+1} - \theta_i) + \delta \sin(\theta_{i-1} - \theta_i), \text{ for } i = 2 \dots N-1, \\ \dot{\theta}_N &= \omega_N + \delta \sin(\theta_{N-1} - \theta_N).\end{aligned}\tag{4.4.1}$$

Here we assume that there is a network of interneurons and motoneurons within each segment that generate periodic bursts, and that the state of each “segmental network” can be tracked by a single phase variable θ_i . We allow a frequency gradient $(\omega_1, \dots, \omega_N)$ along the cord from rostral (head, $i = 1$) to caudal (tail, $i = N$) end, and ascending and descending intersegmental coupling is represented by the terms $\alpha \sin(\theta_{i+1} - \theta_i)$ and $\delta \sin(\theta_{i-1} - \theta_i)$ respectively. We assume that $\alpha, \delta > 0$, so that the coupling tends to pull oscillator phases together: cf. Exercise 27. Wilson [231, §13.2] presents a generalized version of this in the course of a review of models from the segmental level to that of the whole cord, so we only give a brief sketch here.

Provided that the frequency differences $\omega_{i+1} - \omega_i$ are small, and/or coupling strengths α, δ are large, traveling wave solutions can be found for (4.4.1). This is most easily seen by *assuming* a uniform, time-independent phase lag $\theta_{i+1} - \theta_i \stackrel{\text{def}}{=} \gamma$, substituting into (4.4.1), and subtracting the equations pairwise:

$$\begin{aligned}\dot{\theta}_2 - \dot{\theta}_1 &= (\omega_2 - \omega_1) - \alpha[\sin(\gamma) - \sin(\gamma)] + \delta \sin(-\gamma), \dots, \\ \dot{\theta}_{i+1} - \dot{\theta}_i &= (\omega_{i+1} - \omega_i) + \alpha[\sin(\gamma) - \sin(\gamma) + \delta[\sin(-\gamma) - \sin(-\gamma)]] = 0, \dots, \\ \dot{\theta}_N - \dot{\theta}_{N-1} &= (\omega_N - \omega_{N-1}) + \delta[\sin(-\gamma) - \sin(-\gamma)] - \alpha \sin(\gamma).\end{aligned}\tag{4.4.2}$$

If γ is to be constant, then the differences $\dot{\theta}_{i+1} - \dot{\theta}_i$ between time derivatives must all vanish and so we must have

$$(\omega_2 - \omega_1) = \delta \sin \gamma, \quad (\omega_N - \omega_{N-1}) = \alpha \sin \gamma \quad \text{and} \quad \omega_{i+1} - \omega_i = 0 \text{ for } i = 2 \dots N-1,\tag{4.4.3}$$

and evidently $|\omega_2 - \omega_1| < \delta$ and $|\omega_N - \omega_{N-1}| < \alpha$ must also hold for real solutions of the first two conditions of (4.4.3). Moreover, for a uniform wave to propagate from head to tail we require $\gamma < 0$ ($\theta_{i+1}(t)$ lags behind $\theta_i(t)$), so we deduce that the head and tail oscillators must run respectively faster and slower than the interior oscillators, which should all have the same frequency. This is a rather unlikely situation, and in [41] it was also shown that, for the case $\delta = \alpha$, a phase locked solution exists with a uniform *frequency gradient* $\Delta\omega = \omega_{i+1} - \omega_i$ provided that $\Delta\omega < 8\alpha/N^2$. In that case the phase lag $\gamma_i = \theta_{i+1} - \theta_i$ is nonuniform, being greatest in the center of the cord and smallest at the ends, and note that N enters in the condition for phase locking, which becomes more stringent for longer chains.

Actually one should use a double chain of oscillators, since the waves of neural and electromyograph (EMG) activation measured in lamprey spinal cords are in antiphase contralaterally as well as displaying phase lags ipsilaterally: see [42], but as Wilson notes [231, §13.1], “symmetry sub-sampling” techniques may be used to justify analysis of only one side (left or right), provided the model is bilaterally symmetric.

In the introductory paragraph to this section we noted that integrated neuro-mechanical models, including CPGs, muscles, body-limb dynamics and environmental reaction forces, are needed if we are to better understand the rôle of CPGs in producing locomotion. (Indeed, without reaction forces, animals would go nowhere!) Examples of such models for lamprey swimming, of increasing sophistication and realism, can be found in [155, 156, 213].

4.4.2 A CPG model for insects

We now describe some analysis on the phase oscillator reduction of a CPG model for hexapedal locomotion. The model was motivated by experiments of Pearson and Iles [163, 165] on American cockroaches (*Periplaneta americana*), and uses bursting neuron and chemical synapse models of the types described in §3.5 and §4.1. PRCs and averaged coupling functions in terms of phase were computed numerically. For details, see [77, 76]. The architecture of the CPG and motoneuron network is shown in Fig. 74, the caption of which provides some descriptive background.

As Fig. 74(c) indicates, one-way synaptic connections run from the CPG interneurons to the motoneurons, so the basic stepping rhythm is determined by the six CPG units, which may be studied in isolation. But before proceeding, we must explain the insect gait pattern that we wish to model. Cockroaches run over much of their speed range with a double tripod gait, in which the left front, left rear and right middle legs (the L tripod) alternate with the right front, right rear and left middle legs (the R tripod) in providing stance support. Motoneurons activating depressor muscles that provide support and drive the “power stroke” during stance must therefore be alternately active for the L and R tripods: in particular, neighboring legs on the same side (ipsilateral) and across the body (contralateral) should work in antiphase. We augment Pearson’s proposal for three mutually-inhibiting bursting interneurons on each side of the thoracic ganglia (Fig. 74(a)) by similarly linking the front, middle and rear pairs on the two sides by mutual inhibition. In Fig. 74(c) the three cells driving the L tripod are numbered 1, 2, 3 and those driving the R tripod: 4, 5, 6. In fact the hemisegmental ganglia contain multiple active neurons, and the representation of each “leg unit” by a single bursting cell in Fig. 74 is minimal. For example, the hemisegmental units of lamprey described in [231, §13.1] each contain 3 different cell types, cf. [33, 216].

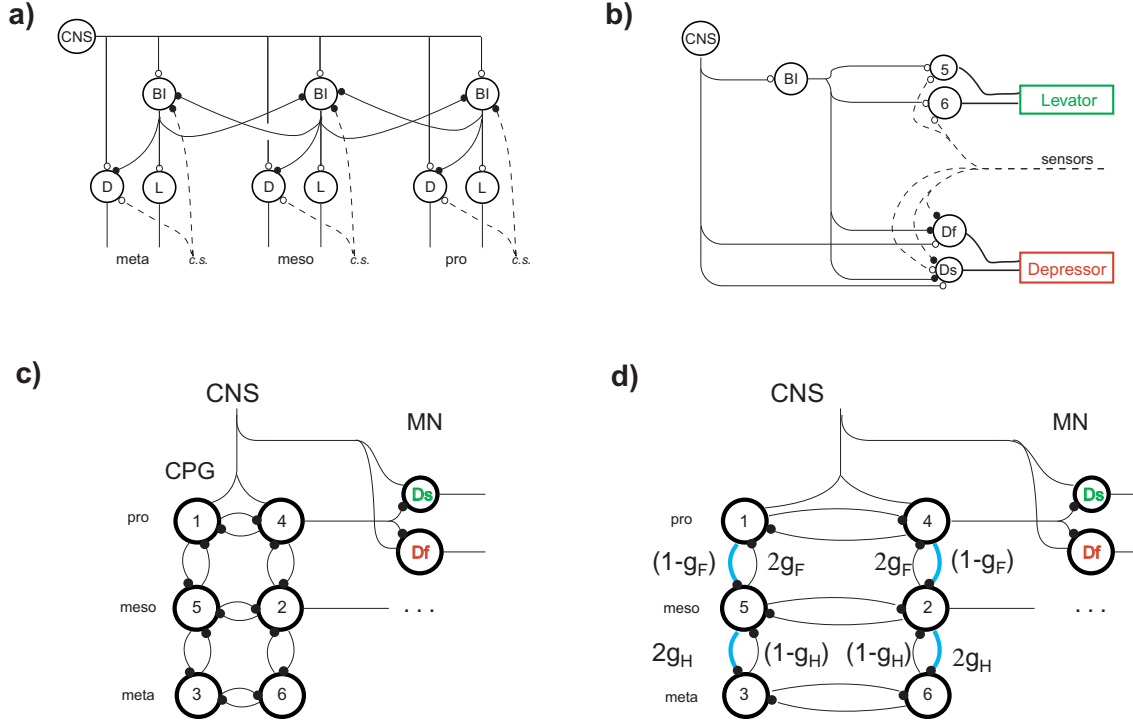


Figure 74: A central pattern generator for hexapedal locomotion. (a) Ipsilateral CPG-motoneuron network connectivity and (b) individual leg depressor and levator circuits in *Periplaneta americana*, showing fast and slow motoneurons D_f , D_s as proposed by Pearson [165]. The central nervous system (CNS) excites bursting interneurons (BI) as well as D_f and D_s , which innervate the leg depressor muscles. Motoneurons 5 and 6 innervate levator muscles and are not modelled here. Sensory feedback (dashed) affects the activity of motoneurons and interneurons. Open circles indicate excitatory coupling, closed circles indicate inhibitory coupling, *c.s.* denotes campaniform sensillae: proprioceptive sensors on the exoskeleton that measure forces in the legs. (c) Network connectivity of the hexapedal model: CPG interneurons are coupled through mutually-inhibiting synapses and fast and slow motoneurons are connected via inhibitory synapses to their corresponding CPG neuron; they are also tonically driven by the CNS. (d) Asymmetric ipsilateral coupling: contralateral coupling strengths \bar{g}_{syn} are equal, and the network of (c) is obtained when $g_F = \bar{g}_{syn}/2 = g_H$. Adapted from [76].

The reduced phase model for the six CPG cells takes the form:

$$\begin{aligned}
\dot{\psi}_1 &= \bar{g}_{\text{syn}} H(\psi_1 - \psi_4) + \bar{g}_{\text{syn}} H(\psi_1 - \psi_5), \\
\dot{\psi}_2 &= \frac{\bar{g}_{\text{syn}}}{2} H(\psi_2 - \psi_4) + \bar{g}_{\text{syn}} H(\psi_2 - \psi_5) + \frac{\bar{g}_{\text{syn}}}{2} H(\psi_2 - \psi_6), \\
\dot{\psi}_3 &= \bar{g}_{\text{syn}} H(\psi_3 - \psi_5) + \bar{g}_{\text{syn}} H(\psi_3 - \psi_6), \\
\dot{\psi}_4 &= \bar{g}_{\text{syn}} H(\psi_4 - \psi_1) + \bar{g}_{\text{syn}} H(\psi_4 - \psi_2), \\
\dot{\psi}_5 &= \frac{\bar{g}_{\text{syn}}}{2} H(\psi_5 - \psi_1) + \bar{g}_{\text{syn}} H(\psi_5 - \psi_2) + \frac{\bar{g}_{\text{syn}}}{2} H(\psi_5 - \psi_3), \\
\dot{\psi}_6 &= \bar{g}_{\text{syn}} H(\psi_6 - \psi_2) + \bar{g}_{\text{syn}} H(\psi_6 - \psi_3).
\end{aligned} \tag{4.4.4}$$

Here the strengths of the inhibitory synapses are chosen so that the net effect on each cell from those connected to it is the same (the middle leg cells 2 and 5 receive inputs from three neighbors; front and hind leg cells from two).

The averaged coupling function H takes the form shown in Fig. 75. Although the PRC is a complicated function with multiple oscillations caused by the burst of spikes, the integral required by the averaging theorem yields a fairly simple function. Indeed, subtraction of $H(-\theta)$ from $H(\theta)$ produces a phase difference function $G(\theta)$ that not only has zeroes at $\theta = 0$ and π , as noted in §4.3.3, but is also odd and remarkably close to a simple sinusoid, as assumed in the earlier phase oscillator model for lamprey CPG [41]. Note that an odd function only appears for symmetric coupling ($G(\theta) = [H_{ji}(\theta) - H_{ji}(-\theta)]$).

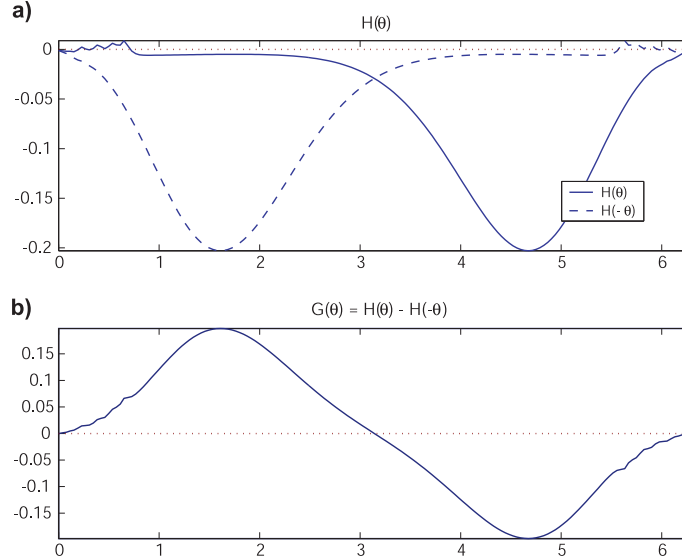


Figure 75: (a) The coupling function $\bar{g}_{\text{syn}} H_{ji}(\theta)$ (solid) for an inhibitory synapse; $\bar{g}_{\text{syn}} H_{ji}(-\theta)$ also shown (dash-dotted). (b) The phase difference coupling function $\bar{g}_{\text{syn}} G(\theta) = \bar{g}_{\text{syn}} [H_{ji}(\theta) - H_{ji}(-\theta)]$. Note that $G(0) = G(\pi) = 0$ and $\bar{g}_{\text{syn}} G'(0) > 0 > \bar{g}_{\text{syn}} G'(\pi)$. From [76].

If we seek L-R tripod solutions of the form $\psi_1 = \psi_2 = \psi_3 \equiv \psi_L(t)$, $\psi_4 = \psi_5 = \psi_6 \equiv \psi_R(t)$, then Eqns. (4.4.4) collapse to the pair of ODEs

$$\dot{\psi}_L = 2\bar{g}_{\text{syn}} H(\psi_L - \psi_R) \text{ and } \dot{\psi}_R = 2\bar{g}_{\text{syn}} H(\psi_R - \psi_L), \tag{4.4.5}$$

and the arguments used in §4.3.3 may be applied to conclude that $\psi_R = \psi_L + \pi$ and $\psi_R = \psi_L$ are fixed points of (4.4.5), *independent* of the precise form of H . For this argument to hold, note

that the sums on the right hand sides of the first three and last three equations of (4.4.4) must be identical when evaluated on the tripod solutions; hence, net inputs to each cell from its synaptic connections must be equal. Also, since for $\bar{g}_{\text{syn}} > 0$ we have $G'(0) > 0 > G'(\pi)$ (Fig. 75(b)), so that we *expect* the in-phase solution to be unstable and the antiphase one to be stable. However, to confirm this in the full six-dimensional phase space of (4.4.4) we must form the 6×6 matrix:

$$\bar{g}_{\text{syn}} \begin{bmatrix} 2H' & 0 & 0 & -H' & -H' & 0 \\ 0 & 2H' & 0 & -H'/2 & -H' & -H'/2 \\ 0 & 0 & 2H' & 0 & -H' & -H' \\ -H' & -H' & 0 & 2H' & 0 & 0 \\ -H'/2 & -H' & -H'/2 & 0 & 2H' & 0 \\ 0 & -H' & -H' & 0 & 0 & 2H' \end{bmatrix}, \quad (4.4.6)$$

where the derivatives H' are evaluated at the appropriate (constant) phase differences π or 0 . The anti-phase tripod solution $\psi_L - \psi_R = \pi$ gives one zero eigenvalue with “equal phase” eigenvector $(1, 1, 1, 1, 1, 1)^T$, and the remaining eigenvalues and eigenvectors are as follows:

$$\begin{aligned} \lambda &= \bar{g}_{\text{syn}} H' : (1, 0, -1, 1, 0, -1)^T, \\ \lambda &= 2\bar{g}_{\text{syn}} H', m = 2 : (1, -1, 1, 0, 0, 0)^T \text{ and } (0, 0, 0, -1, 1, -1)^T, \\ \lambda &= 3\bar{g}_{\text{syn}} H' : (1, 0, -1, -1, 0, 1)^T, \\ \lambda &= 4\bar{g}_{\text{syn}} H' : (1, 1, 1, -1, -1, -1)^T. \end{aligned} \quad (4.4.7)$$

(This can be checked by Matlab.) Since $\bar{g}_{\text{syn}} H'(\pi) < 0$ (Fig. 75(a)), this does prove asymptotic stability with respect to perturbations that disrupt the tripod phase relationships; moreover, the system recovers fastest from perturbations that disrupt the relative phasing of the L and R tripods ($\lambda = 4\bar{g}_{\text{syn}} H'$: last entry of (4.4.7)). Since $\bar{g}_{\text{syn}} H'(0) > 0$ (Fig. 75(a)), the “pronking” gait with all legs in phase ($\psi_L(t) \equiv \psi_R(t)$) is unstable.

Exercise 29. *How many other phase locked solutions of the six ODEs (4.4.4) can you find? [Hint: Seek solutions in which different groups of 3 (and then of 2) phases are equal, e.g. $\psi_1 = \psi_5 = \psi_3 = \psi_L$, etc.]*

Exercise 30. *Consider the CPG network with unequal ipsilateral coupling shown in Fig. 74(d). Can you find a modified double tripod solution with each contralateral pair of cells in antiphase, but with nonzero phase differences between front and middle and middle and hind cells? [Hint: As in §4.4.1 assume that such a solution exists and try to find conditions on the synaptic strengths g_F and g_H consistent with it.]*

This CPG model was created in the absence of information on coupling strengths among different hemisegments, and symmetry assumptions were made for mathematical convenience, allowing the reduction to a pair of tripod oscillators, as in Eqns. (4.4.5). Recent experiments support bilateral symmetry (L-R and R-L contralateral connections have the same strength), but indicate that descending connections are stronger than ascending ones [68], as allowed in Fig. 74(d). Similar rostralcaudal asymmetries have been identified in the lamprey spinal cord [92, 10].

In closing, we note that neuro-mechanical models of insect locomotion have also been constructed that integrate a CPG with muscles and body-limb dynamics [133, 134]. They were developed from much simpler, bipedal models with passive, springy legs [188, 189, 187] by introducing feed-forward

actuation [190], constructing a more realistic, hexapedal geometry driven by agonist-antagonist muscle pairs [132], and finally adding proprioceptive feedback [170]. Their rigid bodies are equipped with jointed, massless legs whose ground contact points remain fixed, and the body moves in the horizontal plane, implying that its dynamics have only 3 degrees of freedom (fore-aft and lateral translations and yawing rotation). In spite of these simplifications, the final, unanalyzable system contains 270 ODEs and simulations of it run too slowly to reveal how its behavior depends on multiple parameters. It has subsequently been reduced to 24 phase oscillators describing motoneurons driving extensors and flexors, coupled to 6 force and moment balance equations describing the Newtonian body dynamics [171]. Linearization around the limit cycles implicit in phase reduction allows one to interpret the contributions of internal CPG coupling and feedback additively, thereby illuminating mechanisms involved in stability and recovery from perturbations.

5 Probabilistic methods and information theory (*≈ 2 weeks*)

In §3 the emphasis was on biophysically-based mechanistic models: sets of deterministic ODEs describing the evolution of membrane voltages, ionic currents and gating variables, and reductions of them. We now turn to empirical and probabilistic models, describing how they can be formulated from experimental data. Specifically, we will examine neural coding and decoding, addressing two questions: 1) What spike train results from a given stimulus? and 2) Given a spike train, what was the stimulus that caused it? To develop (partial) answers, we provide a brief introduction to relevant ideas from probability and information theory, and their implications for efficiency and reliability of neural coding and communication. The main sources for this section are the books of Rieke et al. [179] and Dayan and Abbott [50].

Unlike the continuous membrane potentials $v(t)$ of §3 (but like the integrate and fire model), here we shall describe neural signals as trains of delta functions. Since action potentials are stereotyped and subthreshold fluctuations attenuate on very short scales in axons, the times at which soma voltages peak often suffice to describe the output of a neuron. As in the integrate-and-fire models of §4.2.2 this delta function idealization provides a compact description: for example, a one (or ten) second recording containing fifty spikes has only fifty elements in its data set – the fifty spike times – representing a substantial compression of the membrane potential discretized on a millisecond timescale.

The simplification is appropriate given the questions we now address. Instead of probing physiological mechanisms that generate spikes, we ask how spike trains convey information to the organism about the outside world. Details about mechanisms are largely irrelevant to such a question, so we abstract from the physiology to make progress. However, we must not forget that spike trains come from real neurons, which are not idealized binary (0, 1) communication units, but are composed of ion channels, axons, dendrites, synapses, etc., as described in §3. Indeed, computation of a sort does occur at the cellular level (e.g., when subtle ionic current timescale effects determine whether a given cell reaches threshold and fires a spike), but here we emphasize information contained in and transmitted by (idealized) spike trains themselves.

The encoding/decoding questions posed above can be further refined. The central nervous system receives sensory input in the form of spike trains, which contain visual, auditory, and tactile

information from the eyes, ears, and skin receptors. How is this information conveyed? Given a stimulus, what spike train results? How does the organism extract stimulus information from a series of spike times? How closely do specific neural coding schemes approach to optimal information transmission rates? How robust and reliable are encoding and decoding? We will not completely answer any of these questions, but we will develop tools necessary to address them. We start with statistical and probabilistic tools for spike train analysis, continue to filtering techniques for decoding, and end with an introduction to information theory relevant for the final two questions.

5.1 Stimulus to response maps: Encoding

After introducing idealized models of spike trains and different notions of averaged spike rates, we review basic ideas in probability theory. Equipped with these, we return to spike trains to describe spike-triggered averages (patterns evoked by stereotypical stimuli), and a simple statistical model for spiking

5.1.1 Description of spike trains

As noted above, we consider only spike times in the neural recording, rather than the full membrane potential. Hence, given a train of n spikes over an interval $[0, T]$, our data consists of n spike times, $\{t_i\}$, and the spike train as a function of time becomes:

$$\rho(t) = \sum_{i=1}^n \delta(t - t_i). \quad (5.1.1)$$

Recall that the delta function $\delta(t)$ is a *generalized function*, defined by what it does to continuous functions rather than via its pointwise values. Specifically, $\delta(t) = 0$ for all $t \neq 0$, and while its value is not defined at $t = 0$, it has unit integral $\int_{-\infty}^{+\infty} \delta(t) dt = 1$. Thus, the integral

$$\int_0^T \rho(t) dt = n \quad (5.1.2)$$

counts the number of spikes (n) in the interval $[0, T]$, and when convolved with a continuous function $f(t)$, $\rho(t)$ sifts out the value of that function at the time of the spike: $\int_{-\infty}^{+\infty} \delta(t - t_0) f(t) dt = f(t_0)$. Therefore, from (5.1.1),

$$\int_{-\infty}^{\infty} f(\tau) \rho(t - \tau) d\tau = \sum_{i=1}^n f(t - t_i), \quad (5.1.3)$$

and sums over spikes are interchangeable with integrals with the neural output $\rho(t)$.

Spike sequences are usually stochastic: if the same stimulus is presented multiple times, spike sequences from the same neuron may differ due to internal noise in the system. (Indeed, delta functions cannot exactly line up, when one considers that $\{t_i\}$ has measure zero in $[0, T]$.) Nonetheless, spike trains elicited by the same stimulus under carefully controlled conditions usually “look similar,” and we would like to quantify this similarity using probability and statistics. Small jitter

in spike times does not necessarily pose a problem: microsecond differences often do not register to the organism, and in many cases, *firing rates* rather than individual spike times appear to code for sensory properties. (Echolocation in directional hearing by bats and owls is a stroking counterexample.)

Following the notation in [50] we explore different definitions and meanings of the firing rate. The first is the *spike count rate* r : the average number of spikes per unit time during stimulus presentation. This lacks time resolution, but can be very useful for constant stimuli. The definition uses the delta function property of Eqn. (5.1.2):

$$r = \frac{n}{T} = \frac{1}{T} \int_0^T \rho(\tau) d\tau. \quad (5.1.4)$$

The spike count rate r is useful in that it requires only one stimulus presentation to obtain, but it averages out time variations and is thus useless for treating non-stationary stimuli, unless T is very small. However, we can divide the train into time bins of width Δt , and at a suitable resolution, each bin will contain zero or one spike, resulting in a rate that is piecewise either 0 or $1/\Delta t$. Presenting the same stimulus multiple times to the neuron, we may then derive an average response $r(t)$ (note the use of Roman r to distinguish from $r(t)$). Over fifty presentations, a given bin may have 15 spikes, which gives an instantaneous rate of $15/50\Delta t = 0.3/\Delta t$. The values are still discrete (multiples of $1/\Delta t$ divided by the number of presentations), but they offer much better resolution than 0 or 1. There is a tradeoff between *temporal resolution* (making Δt small) and *rate resolution* (possible values for $r(t)$). Letting angle brackets $\langle \rangle$ denote averaging over trials, we define the *time-dependent firing rate* (or simply the *firing rate*) as:

$$r(t) = \frac{1}{\Delta t} \int_t^{t+\Delta t} \langle \rho(\tau) \rangle d\tau. \quad (5.1.5)$$

The limit $\Delta t \rightarrow 0$ is not taken since that would require the number of trials to approach infinity to avoid the rate resolution problem. The probability of finding a spike in a window of size Δt , starting at time t , is then $r(t)\Delta t$.

Finally, we define the *average firing rate* $\langle r \rangle$ as the average number of spikes in the spike train averaged over trials. It has the same drawbacks as r , and can be defined in three equivalent ways:

$$\langle r \rangle = \frac{\langle n \rangle}{T} = \frac{1}{T} \int_0^T \langle \rho(t) \rangle dt = \frac{1}{T} \int_0^T r(t) dt. \quad (5.1.6)$$

Note that for the last equality, the limit $\Delta t \rightarrow 0$ must be taken, for which $r(t) \rightarrow \langle \rho(t) \rangle$ in (5.1.5).

Firing rates may be determined by recording spike trains for (identical) stimulus presentations and summing them aligned with $t = 0$ corresponding to stimulus onset. The summed record is divided into bins of width Δt and the total number of spikes in each bin across all trials is counted and divided by Δt to get the firing rate over that bin. The result is then plotted as a Post (or Peri) Stimulus Time Histogram (PSTH): Fig. 76.

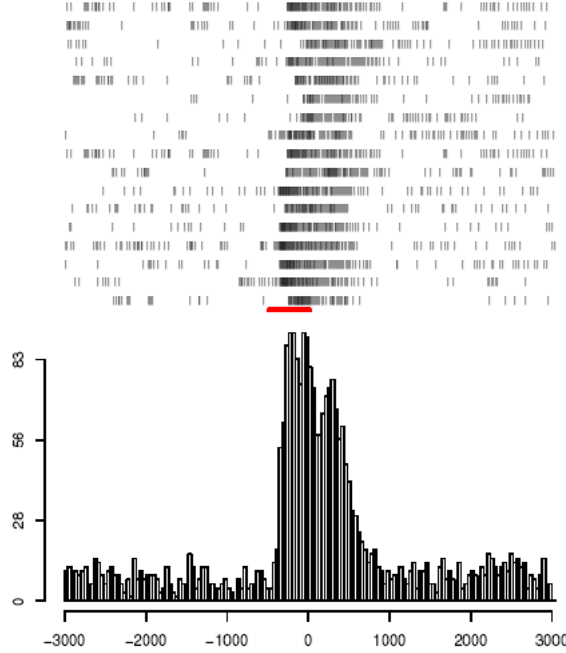


Figure 76: Construction of the Peri Stimulus Time Histogram (PSTH) from 17 0.5 second stimulus presentations. Individual spike trains are shown on the raster plot above and the PSTH below with spike rates in Hz. Here the stimulus (red bar below raster plot) is centered on $t = 0$, not starting at $t = 0$, as is usual. www.biomedicale.univ-paris5.fr

5.1.2 A primer on probability theory

We assume familiarity with basic concepts in probability, but we review a few key ideas here. For additional background see a text such as [183]. In a discrete probability distribution, there are N possible outcomes $\{X_i\}_{i=1}^N$, each with probability $P[X_i] = p_i$. Each $p_i \geq 0$, and $\sum_{i=1}^N p_i = 1$. A continuous distribution defines a probability density as a function on some connected domain. For instance, if an event is known to occur between time $t = 0$ and T , its probability can be expressed as a positive function, not necessarily continuous, $p(t) : [0, T] \rightarrow \mathbb{R}^+$, where $\int_0^T p(t) dt = 1$. The probability that the event occurs between times a and b (with $0 \leq a < b \leq T$) is the integral of $p(t)$ over $[a, b]$. A discrete distribution can be represented as a sum of weighted delta functions $p_i \delta(X_i)$.

The *mean* or *expected value* μ of the distribution is the average outcome weighted by the probabilities. For discrete distributions

$$\mu = \mathbb{E}[X] = \sum_{i=1}^N X_i p_i, \quad (5.1.7)$$

and for a continuous distribution on a domain Ω :

$$\mu = \mathbb{E}[X] = \int_{\Omega} x p(x) dx. \quad (5.1.8)$$

The *variance* of a distribution $\text{Var}(X)$ or σ^2 , with σ the *standard deviation*, is a measure of

how much the distribution “spreads” about the mean. For a discrete distribution

$$\text{Var}(X) = \sigma^2 = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2 = \sum_{i=1}^N X_i^2 p_i - \mu^2, \quad (5.1.9)$$

and in the continuous case:

$$\sigma^2 = \int_{\Omega} (x - \mu)^2 p(x) dx. \quad (5.1.10)$$

Certain distributions appear frequently in nature and will be very important in this section. The most common is the *Gaussian*, or *normal distribution*, a continuous distribution with density

$$p(x) = N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right], \quad (5.1.11)$$

having mean μ and standard deviation σ . Normal distributions are important in applications because the average of n independent, identically-distributed (i.i.d.) random variables, each with mean μ and variance σ^2 , is normally distributed with mean μ and variance σ^2/n , by the central limit theorem [183, p.404]. Since neural events often result from multiple independent upstream events, it is not surprising to see Gaussians in neuroscience.

The exponential distribution, which is often used to model the distribution of times until a single event occurs (e.g., failure of a lightbulb), is another useful example. This distribution is given by

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{for } x \geq 0, \\ 0, & \text{for } x < 0. \end{cases} \quad (5.1.12)$$

The probability that the waiting time will be longer than T , given a time constant λ , is calculated by summing the probability in the tail of the distribution for $x > T$ as follows:

$$P[t > T] = \int_T^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda T}. \quad (5.1.13)$$

Interspike intervals are often modeled by repeated drawings from (5.1.12) with a suitable time constant λ . We leave computation of the mean and variance of this distribution as Exercise 32 below. A double-sided version of the exponential distribution will appear in Exercise 34 of §5.2.

The *Poisson distribution* is an important discrete distribution, in which X takes integer values $i = 0, 1, 2, \dots$ and

$$p_i = P[X = i] = \frac{\lambda^i e^{-\lambda}}{i!}. \quad (5.1.14)$$

Here the parameter λ is both the mean and the variance, as can easily be checked using (5.1.7) and (5.1.9). The Poisson distribution arises from the sum of $n \gg 1$ trials, each with small probability p of the event occurring, so that $np = \lambda$ is neither small nor large. Thus you can easily see how λ will end up as the mean for that distribution. It provides a good description of the total occurrences of an event, if each is independent of the others. In §5.1.4 we will use it in a spike-generating model.

Exercise 31. Verify that the mean and variance for (5.1.14) are both λ , and that they are μ and σ^2 for (5.1.11). [Hint: You will need to let $i \rightarrow \infty$ for the Poisson distribution.]

Two quantities are commonly used to describe the spread of probability distributions. The *coefficient of variation* is the ratio $CV = \sigma/\mu$ of the standard deviation to the mean. The *Fano factor* is the ratio $F = \sigma^2/\mu$ of variance to mean. For Poisson distributions the Fano factor is 1, and determining it for spike counts shows how close their distributions are to Poisson [50, p.32].

Multiple random variables can be related to the same distribution of outcomes. For instance, let X be the random variable of the ordered pair of results from rolling a red and a white die. Another random variable Y could represent the sum of the two dice. Possible values of X are $(1,1), (2,3), (3,2), \dots$, each with probability $1/36$ for fair dice, and possible values of Y are the integers 2 through 12, with varying probabilities. Suppose we know the value of Y for a given trial. What do we know about X ? $Y = 2$ implies that $X = (1,1)$, but if $Y = 7$ there are six possibilities for X : $[(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)]$, each with probability $1/6$. So we may define a new distribution describing the probability of X , once the value of Y is known. This is called the *conditional probability of X given Y* , and is usually written $P[X|Y]$. It will be very important when we describe the distribution of spike times $\{t_i\}$ given the stimulus $s(t)$ as $P[\{t_i\}|s(t)]$ for stimulus encoding and its “dual” $P[s(t)|\{t_i\}]$ for decoding.

The *joint distribution* $P[X, Y]$ of two random variables X and Y (or the joint density $p(x, y)$ for continuous variables) describes the probabilities of different combinations of X and Y . The *marginal distribution* of Y can be obtained from the joint probability by summing or integrating over X . In the continuous formulation, this is expressed as

$$p(y) = \int_{-\infty}^{\infty} p(x, y) dx. \quad (5.1.15)$$

The conditional distribution of X given Y can then be obtained from the joint distribution by dividing the joint probability by the probability for Y . For instance, if the probability that $X = Y = 5$ is .1 and the probability that $Y = 5$ is .5, then the probability that $X = 5$ given that $Y = 5$ is

$$P[X = 5|Y = 5] = \frac{P[X = 5, Y = 5]}{P[Y = 5]} = \frac{0.1}{0.5} = 0.2.$$

So given the joint distribution and the distribution of Y , we get the conditional distribution of X given Y . The reverse is also true, and in general we have

$$P[X|Y] P[Y] = P[X, Y] = P[Y|X] P[X]. \quad (5.1.16)$$

In the neural spike train context, experiments usually determine only $P[\{t_i\}|s(t)]$ or $P[s(t)|\{t_i\}]$, but we typically require both distributions. Fortunately, from (5.1.16) we can get the joint distribution $P[\{t_i\}, s(t)]$ from both $P[\{t_i\}|s(t)]P[s(t)]$ and from $P[s(t)|\{t_i\}]P[\{t_i\}]$. Setting the two representations equal yields Bayes’ rule [13]:

$$P[s(t)|\{t_i\}] = P[\{t_i\}|s(t)] \frac{P[s(t)]}{P[\{t_i\}]}. \quad (5.1.17)$$

We will return to this when discussing encoding and decoding.

Finally, two random variables X and Y are *independent* if $P[X, Y] = P[X] P[Y]$. Hence $P[X|Y] = P[X]$: knowing Y tells us nothing about X , and vice versa. For continuous distributions, independence holds if and only if $p(x, y) = p(x)p(y)$: the joint density is equal to the product of the individual densities.

Exercise 32. Find the mean and variance of the exponential distribution for time constant λ . Also, given that the event has not happened by time T_0 , what is the conditional probability that the event will happen after time $T_0 + T$? Why is this interesting?

5.1.3 Spike triggered averages

We now develop quantitative analyses for the generation of spike trains, given stimuli. We assume that spike trains have been recorded in response to a stereotypical stimulus $s(t)$ repeated many times, and the time-dependent firing rate $r(t)$ has been obtained by averaging over trials. We define the *spike-triggered average stimulus* (STA) $C(\tau)$ as the average value of the stimulus at time τ before a spike [50]. We average over all spikes, both within and across trials to obtain

$$C(\tau) = \left\langle \frac{1}{n} \sum_{i=1}^n s(t_i - \tau) \right\rangle \approx \frac{1}{\langle n \rangle} \left\langle \sum_{i=1}^n s(t_i - \tau) \right\rangle. \quad (5.1.18)$$

We may take $\langle n \rangle$ out of the trial average if n is large enough that $n \approx \langle n \rangle$. $C(\tau)$ should approach the mean of the stimulus at large τ , since outside of a certain correlation time, we do not expect the stimulus to have anything to do with a spike. Thus, the expected stimulus value distant from spikes is just the value with no information at all: the global average. Using the notation developed in §5.1.1, we note that

$$\int_0^T \langle \rho(t) \rangle s(t - \tau) dt = \left\langle \int_0^T \rho(t) s(t - \tau) dt \right\rangle = \left\langle \sum_{i=1}^n s(t_i - \tau) \right\rangle,$$

where the first step is possible because the same stimulus is applied in every trial, and the second follows from (5.1.1). Hence we may write (5.1.18) as

$$C(\tau) = \frac{1}{\langle n \rangle} \int_0^T \langle \rho(t) \rangle s(t - \tau) dt = \frac{1}{\langle n \rangle} \int_0^T r(t) s(t - \tau) dt, \quad (5.1.19)$$

the final step uses $\lim_{\Delta t \rightarrow 0} r(t) = \langle \rho(t) \rangle$, as in (5.1.6).

The expression (5.1.19) resembles the correlation of the firing rate and stimulus, but it differs in an important way. Correlations show how functions are related at different times. The firing rate-stimulus correlation function $Q_{rs}(\tau)$ is

$$Q_{rs}(\tau) = \frac{1}{T} \int_0^T r(t) s(t + \tau) dt, \quad (5.1.20)$$

and it shows how the firing rate at time t is related to the stimulus at $t + \tau$ on average, as a function of τ . From (5.1.19-5.1.20), the spike triggered average $C(\tau)$ is $Q_{rs}(-\tau)$, normalized by $\langle r \rangle = \langle n \rangle / T$. The sign of τ differs in the two definitions, and there is a normalization factor, but the spike triggered average is often called a *reverse correlation function*. Other names noted in [179] include *first Wiener kernel*, *mean effective stimulus*, and *triggered correlation function*. See Fig. 77 for examples of STAs for retinal ganglion cells. The STA $C(\tau)$ is conventionally plotted with τ increasing to the left to emphasize that it represents the stimulus preceding the spike: see Fig. 79 below.

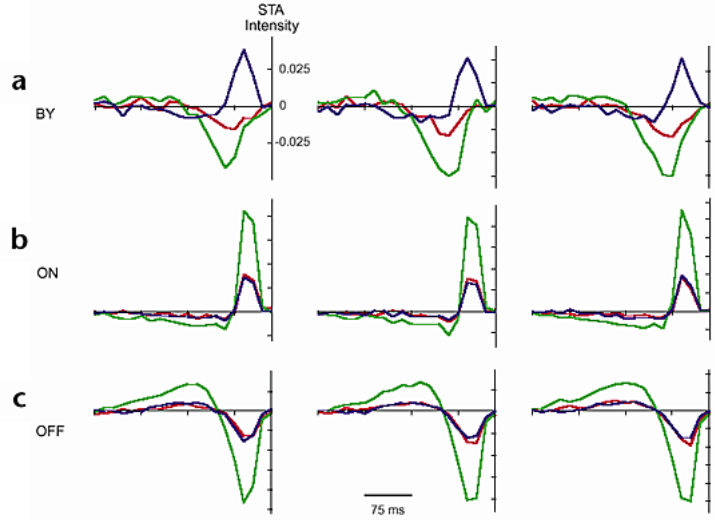


Figure 77: Spike-triggered averages $C(\tau)$ for nine different macaque ganglion cells in the visual system for red, green, and blue color stimuli. Row a shows three different blue-yellow cells, which are seen to respond on average to an increase in blue or a decrease in green or red intensities. Row b shows three different ON cells, and row c shows three different OFF cells. Note that $C(\tau)$ is conventionally plotted with τ increasing to the left. From *www.nature.com Chichilnisky and Baylor, 1999*

Wiener kernels characterize nonlinear systems in terms of certain filters. They resemble Taylor series, with functions rather than real or complex numbers as the domain variables. Wiener showed that the output $y(t) = F[x(t)]$ of a nonlinear system with input $x(t)$ can be written as

$$y(t) = G_0 + G_1[x(t)] + G_2[x(t)] + G_3[x(t)] + \cdots, \quad (5.1.21)$$

with $G_0 = g_0$ a constant term, and G_1 defined as

$$G_1[x(t)] = \int_0^\infty g_1(\tau_1)x(t - \tau_1) d\tau_1. \quad (5.1.22)$$

The function $g_1(\tau_1)$ is called the *first Wiener kernel*; it is analogous to the first derivative in the linear term of Taylor series. Higher order terms have kernels that depend on higher-order cross-correlations, e.g., the average value of the stimulus at times τ_1 from one spike and τ_2 from another. Given enough terms, Wiener series provide good approximations of many nonlinear systems. The simplest stimulus to firing rate encoder, then, is just the first Wiener kernel, which Rieke et al. [179, §A.3, pp 292-295] show is proportional to the spike triggered average. In fact we have

$$g_1(\tau) = \frac{1}{S_x} \langle \int_0^T y(t)x(t - \tau) dt \rangle, \quad (5.1.23)$$

where $\langle \cdot \rangle$ denotes averaging over an ensemble of input-output pairs and S_x is a normalization factor (the power spectrum level).

In summary, we may start with a correlation of firing rate and stimulus, or separately create a filter to approximate the firing rate given the stimulus. Up to a normalizing factor, the filter and

the correlation function are identical:

$$r(t) = G_1[s(t)] = \int_0^\infty g_1(\tau) s(t - \tau) d\tau, \quad (5.1.24)$$

where

$$g_1(\tau) = \frac{1}{S_x} \left\langle \frac{1}{T} \int_0^T \rho(t) s(t - \tau) dt \right\rangle = \frac{1}{S_x} \left\langle \frac{1}{T} \sum_{i=1}^n s(t_i - \tau) \right\rangle, \quad (5.1.25)$$

and $\langle \cdot \rangle$ denotes averaging over an ensemble of stimulus presentations.

This is a very simple model for obtaining the firing rate from stimulus, and in most cases it is far from adequate. Many neural systems exhibit adaptation, as explored earlier, which may happen on slow timescales. The Wiener series has significant difficulties in accounting for adaptation. Good encoding descriptions, even without adaptation, often require higher-order terms. Here we only present simple models that introduce relevant techniques but are certainly not at the research frontier.

5.1.4 A Poisson model for firing statistics

Given the firing rate $r(t)$ corresponding to a stimulus $s(t)$, can we now create the “dictionary” of conditional probabilities of spike trains given the stimulus? Unfortunately, no. There are so many possible spike sequences that correspond to a given approximation of $r(t)$, which itself depends on the choice of Δt , that it is impossible to gather enough trials to determine the full dictionary, even when we have a confident estimate of the firing rate. Moreover, if the presence of one spike alters spike generation in its vicinity through the refractory period or other factors, the firing rate cannot account for this. We must make additional assumptions to obtain a complete encoding model.

Here we describe a simple model of spike generation given the firing rate $r(t)$, which works well in several cases. The major assumption is that spike times are mutually independent. Given the refractory period τ_{ref} , this assumption is not strictly valid for single neurons, since a spike is less likely immediately following another spike, but if τ_{ref} is short compared to the average interspike interval, the model works reasonably well.

For ease of analysis, we start by assuming that $r(t) = r = \text{const}$. If independent spikes arrive at a constant rate, then all trains of n spikes occurring in a given time interval $[0, T]$ have equal probability. We will first find the probability $P[\{t_i\}_{i=1}^n]$ that n spikes occur at specified times $0 < t_1 < t_2 < \dots < t_n < T$, and that no other spikes occur in $[0, T]$. Then we will compute the probability $P_T[n]$ of observing exactly n spikes *anywhere* in the interval $[0, T]$.

We divide the interval into M bins of width $\Delta t = T/M$ and assume that Δt is small enough that not more than one spike falls in each bin (e.g. choose $\Delta t < \tau_{ref}$); in any case, we will eventually take the limit $\Delta t \rightarrow 0$. The probability of a spike occurring in any given bin is simply $r\Delta t$, and, since spike times are assumed to be independent, the joint probability of n spikes occurring in n given bins is therefore $(r\Delta t)^n$. We also require that no spikes occur in the remaining $M - n$ bins, and again appealing to independence, we obtain $(1 - r\Delta t)^{M-n}$, since $(1 - r\Delta t)$ is the probability of not having a spike in a bin. Taking the product of these independent events, we obtain

$$P[\{t_i\}_{i=1}^n] = (r\Delta t)^n (1 - r\Delta t)^{M-n}. \quad (5.1.26)$$

This expression may be simplified by approximating it as $\Delta t \rightarrow 0$ (which implies that $M = T/\Delta t \rightarrow \infty$), while n remains fixed. Thus, $(M - n) \rightarrow M$ and we may write the second factor as

$$\lim_{\Delta t \rightarrow 0} (1 - r\Delta t)^{(M-n)} = \lim_{\Delta t \rightarrow 0} (1 - r\Delta t)^{(T/\Delta t)} = \lim_{\Delta t \rightarrow 0} \left[(1 - r\Delta t)^{-(1/r\Delta t)} \right]^{-(rT)} = e^{-rT}; \quad (5.1.27)$$

in the last step we use the fact that $\lim_{\epsilon \rightarrow 0} (1 + \epsilon)^{1/\epsilon} = e$. This limit provides a good approximation for Δt small, and so, returning to (5.1.26) we have the probability that one spike occurs in each of n specified disjoint bins of size Δt , and that no spikes occur in the remaining $(M - n)$ bins in $[0, T]$:

$$P[\{t_i\}_{i=1}^n] \approx r^n e^{-rT} (\Delta t)^n. \quad (5.1.28)$$

Dividing (5.1.28) by the n -dimensional bin volume, we obtain the probability density function:

$$p[\{t_i\}_{i=1}^n] = r^n e^{-rT}. \quad (5.1.29)$$

To compute the probability $P_T[n]$ that n spikes occur anywhere in the interval $[0, T]$, we return to (5.1.26), and multiply $P[\{t_i\}_{i=1}^n]$ by the number of ways that n spikes can be distributed among M bins, with at most one spike per bin. This is the binomial coefficient M choose n : $\binom{M}{n} = \frac{M!}{(M-n)!n!}$, which gives

$$P_T[n] = \lim_{\Delta t \rightarrow 0} \left[\frac{M!}{(M-n)!n!} (r\Delta t)^n (1 - r\Delta t)^{(M-n)} \right]. \quad (5.1.30)$$

Using the limiting behavior noted above, we deduce that $M!/(M-n)! = \prod_{j=0}^{n-1} (M-j) \approx M^n = (T/\Delta t)^n$ and thus, again appealing to (5.1.27) to simplify the third term in (5.1.30), we obtain

$$P_T[n] = \frac{1}{n!} \left(\frac{T}{\Delta t} \right)^n (r\Delta t)^n \exp(-rT) = \frac{(rT)^n}{n!} \exp(-rT). \quad (5.1.31)$$

$P_T[n]$ is a Poisson distribution with parameter $\lambda = rT$ (cf. Eqn. (5.1.14)), which makes intuitive sense, since the mean number of spikes in the interval $[0, T]$ is the firing rate multiplied by the duration: rT . See Fig. 78 for examples. Note that, using (5.1.31), we may rewrite (5.1.28) as

$$P[\{t_i\}_{i=1}^n] = n! P_T[n] \left(\frac{\Delta t}{T} \right)^n. \quad (5.1.32)$$

The analysis is more complicated when the firing rate $r(t)$ varies, but we can still calculate $P[\{t_i\}_{i=1}^n]$. We again divide the interval $[0, T]$ into M bins of size Δt , and observe that the probability of observing a spike in a given bin containing t_i approaches $r(t_i)\Delta t$ as $\Delta t \rightarrow 0$. Now let M_i denote the bin containing the i th spike ($t_i \in [M_i\Delta t, M_{i+1}\Delta t)$) and consider the $M_{i,i+1} = M_{i+1} - M_i - 1$ bins in the interspike interval (t_i, t_{i+1}) . The probability of not having a spike in any of these is $[1 - r(t_i + j\Delta t)\Delta t]$, and so the joint probability of no spikes occurring in this interval is

$$P[\text{no spikes in } (t_i, t_{i+1})] = \prod_{j=1}^{M_{i,i+1}} [1 - r(t_i + j\Delta t)\Delta t]. \quad (5.1.33)$$

For convenience we transform the product to a sum by taking logarithms, and use the fact that $\ln(1 + \epsilon) = \epsilon + \mathcal{O}(\epsilon^2)$:

$$\ln(P[\text{no spikes in } (t_i, t_{i+1})]) = \sum_{j=1}^{M_{i,i+1}} \ln[1 - r(t_i + j\Delta t)\Delta t] = - \sum_{j=1}^{M_{i,i+1}} r(t_i + j\Delta t)\Delta t + \mathcal{O}(\Delta t^2). \quad (5.1.34)$$

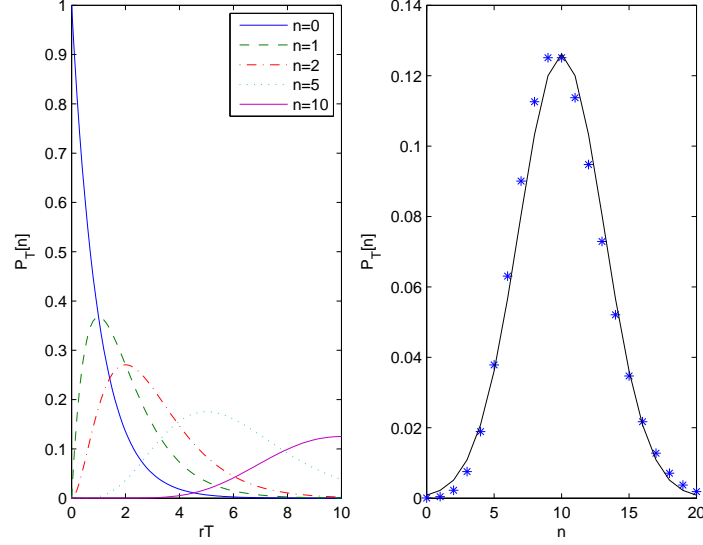


Figure 78: Left: probability that a constant rate Poisson process generates exactly n spikes within time T given rate r , plotted as a function of rT from Eqn. (5.1.31) for various n ; note $n = 0$ gives an exponential distribution. Right: probability of finding n spikes given $rT = 10$, plotted as a function of n for the Poisson process (stars), compared to a Gaussian with $\mu = \sigma^2 = 10$ (solid).

As $\Delta t \rightarrow 0$ this approximation becomes exact and the Riemann sum of (5.1.34) turns into an integral:

$$\ln(P[\text{no spikes in } (t_i, t_{i+1})]) = - \int_{t_i}^{t_{i+1}} r(t) dt. \quad (5.1.35)$$

Exponentiating (5.1.35) and using our earlier observation, we find that the probability of observing a spike in the bin containing t_i and no spikes in the succeeding interval (t_i, t_{i+1}) is well approximated by

$$r(t_i) \Delta t \exp \left(- \int_{t_i}^{t_{i+1}} r(t) dt \right). \quad (5.1.36)$$

Taking products over all spike times and spikeless intervals, we have

$$P[\{t_i\}_{i=1}^n] \approx \exp \left(- \int_0^{t_1} r(t) dt \right) \prod_{i=1}^n \left[r(t_i) \Delta t \exp \left(- \int_{t_i}^{t_{i+1}} r(t) dt \right) \right] \exp \left(- \int_{t_n}^T r(t) dt \right), \quad (5.1.37)$$

and combining all the exponentials this simplifies to

$$P[\{t_i\}_{i=1}^n] \approx \prod_{i=1}^n r(t_i) \exp \left(- \int_0^T r(t) dt \right) (\Delta t)^n, \quad (5.1.38)$$

with the associated probability density function

$$p[\{t_i\}_{i=1}^n] = \prod_{i=1}^n r(t_i) \exp \left(- \int_0^T r(t) dt \right). \quad (5.1.39)$$

In the case $r(t) \equiv r = \text{const.}$, Eqns. (5.1.38) and (5.1.39) reduce to (5.1.28) and (5.1.29), as expected. However, in this case we cannot compute an analog of $P_T[n]$, since the probability of observing spikes varies over the interval $[0, T]$.

Although Eqn. (5.1.39) is not as simple as the constant rate density (5.1.29), it still provides a compact description of spike train generation given a firing rate. It is easy to construct simulations that give sample spike trains given a Poisson distribution, so we now have sample encoding. We can also simulate by interspike intervals: at each spike, choosing the next interspike interval from the appropriate distribution. For the constant rate Poisson process, the distribution of interspike times is an exponential distribution with the form that you will find in Exercise 33. We can also modify the interspike distribution to account for the refractory period, which results in a gamma distribution. For more details see [50, p.33].

Exercise 33. *Show that the distribution of interspike intervals for a constant rate Poisson spike model is an exponential distribution.*

Many types of sensory neurons are well-modeled by Poisson spike processes, which provide good approximations over durations T significantly larger than the refractory period but short enough that low frequency noise in the neural circuit does not overwhelm the distribution. For instance, the encoding of pure tones in primary auditory cortex can be described by a Poisson process over a wide range of stimulus amplitudes [179, p.53].

5.2 Response to stimulus maps: Decoding

While encoding can require complicated nonlinear models, the linear methods developed above often suffice for the related decoding problem: estimating the stimulus $s(t)$ from a spike train $\{t_i\}$. Decoding plays a key role as animals make behavioral choices given neural input about their environment. We remark that reconstruction of $s(t)$ from $\{t_i\}$ may not be the exact problem solved by the organism. For instance, as a fly navigates visually it receives angular motion signals from the H1 neuron (a system extensively studied by Bialek and others [179]). The fly does not necessarily compute angular velocity and then create a flight torque to compensate: torque may be directly computed from the H1 signal. However, the latter process also produces an analog signal related to the stimulus, so it would have a similar structure.

Three issues must be addressed before we develop a simple decoding model. The first is the lack of a well-defined firing rate, on which we based most of the analysis of §5.1. The second is dependence on the stimulus distribution, as in Bayes' rule. The third concern is causality in the decoding problem.

We dealt with the problem of stochasticity in encoding by finding a time-dependent firing rate $r(t)$ (Eqn. (5.1.5)) given multiple stimulus presentations, instead of focusing on single instances of spike trains $\{t_i\}$. Alternatively, we can give up time dependence and compute an average firing rate for a stationary stimulus by assuming a constant rate for time T (the r of Eqn. (5.1.4)). In decoding, the organism has neither luxury because natural stimuli are typically non-stationary (varying on similar time scales to interspike intervals), and demand response after a single viewing.

In many systems, such as photon counting in ganglion cells, spatial memory in rat hippocampus, visual cortical neurons in monkeys, bat auditory cortex during echolocation, speech comprehension in auditory cortex, and animal calls in crickets and frogs, neurons fire at most one spike before the stimulus changes significantly [179, pp.57-59]. In the monkey visual system, fixations last perhaps 100 ms, and cortical neurons fire at rates of 10-50 Hz, so during a given fixation, each neuron might fire 1-5 spikes. One spike cannot determine a rate, and decisions based on the stimulus must be made rapidly. A moth cannot take time to estimate rates and compute evasive paths when a bat is closing in on it. It is unlikely that rate codes are used in such sensory systems.

In some animals many neurons respond to a particular stimulus, allowing simultaneous ensemble averaging to obtain $r(t)$, but this *population model* is also problematic. Such dedicated neural populations are typically small for invertebrates, and in some (e.g., H1 in fly) they are a singleton. Second, even in large populations, rate calculation can be noisy, because averaging requires independence to gain accuracy, and connected neurons may be highly correlated through synaptic excitation and inhibition. Finally, there is a balance between accurately reconstructing stimuli and having neurons do other tasks. How much benefit do multiple neurons give, for the cost of having them compute in parallel instead of something else, or even not existing? Again, we see that individual spike times rather than firing rates can be crucial to decoding.

It is often easier to determine the encoding distribution $P[\{t_i\}|s(t)]$ than the decoding distribution $P[s(t)|\{t_i\}]$ from experiments, unless the train in the latter contains only one or two spikes. Does a full description of $P[\{t_i\}|s(t)]$ suffice to determine $P[s(t)|\{t_i\}]$? The two conditional distributions resemble the sections of a dictionary for translation between English and another language, say Greek. In encoding we use the English-to-Greek section, in decoding Greek-to-English¹¹. Does the former imply the latter? To answer this we recall Bayes' rule (5.1.17):

$$P[s(t)|\{t_i\}] = P[\{t_i\}|s(t)] \frac{P[s(t)]}{P[\{t_i\}]} . \quad (5.2.1)$$

Here $P[\{t_i\}]$ acts as a normalization factor, but the stimulus distribution $P[s(t)]$ establishes the *context*. Without this crucial information, it is impossible to decode a signal, even with full knowledge of the encoding dictionary. This is echoed in both our language analogy and in natural signals.

An example with a very simple encoding scheme is given by Rieke et al. [179]. Suppose that a neuron responds to a stimulus s by producing output x with noise η :

$$x = s + \eta . \quad (5.2.2)$$

Both stimulus and noise are characterized by distributions. There is a distribution for the noise and for the stimulus, but other than the noise, the encoding scheme is linear, and the encoding dictionary $P[x|s]$ is therefore very simple:

$$P[x|s] = P[\eta = x - s] : \quad (5.2.3)$$

the probability of output x given s is exactly the probability that noise η is equal to the difference $x - s$. Assuming that noise samples are drawn from a distribution P_{noi} , we may calculate the decoding distribution from Bayes' rule (5.2.1):

$$P[s|x] = \frac{P[x|s]P[s]}{P[x]} = \frac{1}{P[x]} P_{noi}[\eta = x - s] P[s] . \quad (5.2.4)$$

¹¹If you are Greek, apply the mapping English \leftrightarrow Greek.

How do we determine the best estimate of s given x ? A frequent measure of best is the χ^2 or mean square error: the average of the squared difference between estimated and actual values

$$\chi^2(s, s_{est}) = \langle |s(t) - s_{est}(t)|^2 \rangle. \quad (5.2.5)$$

It can be proved that χ^2 is minimized by an estimate equal to the conditional mean, which is similar to the mean from (5.1.8), but using conditional probability:

$$s_{est} = \int s p[s|x] ds. \quad (5.2.6)$$

Noise is usually modeled as a zero-mean Gaussian with variance $\langle \eta^2 \rangle$:

$$P_{noi}(\eta) = \frac{1}{\sqrt{2\pi\langle\eta^2\rangle}} \exp\left[-\frac{\eta^2}{2\langle\eta^2\rangle}\right], \quad (5.2.7)$$

and substituting (5.2.7) into (5.2.4), we obtain

$$P[s|x] = \frac{1}{\sqrt{2\pi\langle\eta^2\rangle}} \exp\left[-\frac{(s-x)^2}{2\langle\eta^2\rangle}\right] \frac{P[s]}{P[x]}. \quad (5.2.8)$$

To proceed we need the distribution of stimuli $P[s]$. If this is also Gaussian, with mean 0 and variance $\langle s^2 \rangle$, then

$$P[s|x] = \frac{1}{P[x]} \frac{1}{2\pi\sqrt{\langle s^2 \rangle \langle \eta^2 \rangle}} \exp\left[-\frac{s^2}{2\langle s^2 \rangle}\right] \exp\left[-\frac{(s-x)^2}{2\langle \eta^2 \rangle}\right]. \quad (5.2.9)$$

Combining terms and defining a new normalization factor, we get

$$P[s|x] = \frac{1}{Z(x)} \exp\left[-\frac{s^2}{2}\left(\frac{1}{\langle s^2 \rangle} + \frac{1}{\langle \eta^2 \rangle}\right) + s\left(\frac{x}{\langle \eta^2 \rangle}\right)\right], \quad (5.2.10)$$

where $Z(x, \langle s^2 \rangle, \langle \eta^2 \rangle) = P[x] 2\pi\sqrt{\langle s^2 \rangle \langle \eta^2 \rangle} \exp(x^2/2\langle \eta^2 \rangle)$ is fixed, given x ; in particular, the variance $\langle s^2 \rangle$ depends on the distribution from which s is drawn, but does not change once that distribution is chosen. Moreover, $P[s|x]$ is a Gaussian distribution and for Gaussians, the mean is also the value at which the density is maximum, so we may set $\partial P[s|x]/\partial s = 0$ to obtain the condition

$$-s\left(\frac{1}{\langle s^2 \rangle} + \frac{1}{\langle \eta^2 \rangle}\right) + \left(\frac{x}{\langle \eta^2 \rangle}\right) = 0. \quad (5.2.11)$$

This implies that our best estimate is

$$s_{est} = x\left(\frac{\frac{1}{\langle \eta^2 \rangle}}{\frac{1}{\langle s^2 \rangle} + \frac{1}{\langle \eta^2 \rangle}}\right) = x\left(\frac{SNR}{SNR + 1}\right) \stackrel{\text{def}}{=} K_1(SNR)x, \quad (5.2.12)$$

where $SNR = \langle s^2 \rangle / \langle \eta^2 \rangle$ denotes the *signal to noise ratio*. The decoder is evidently also linear, with gain K_1 a function of SNR . The greater the noise, the lower the stimulus estimate will be, and as $SNR \rightarrow 0$, $s_{est} \rightarrow 0$. So far, so good; but if we replace the Gaussian stimulus distribution by the two-sided exponential distribution, $P[s] = (1/2s_0) \exp(-|s|/s_0)$, for example, we get a nonlinear decoder.

Exercise 34. *Given the two-sided exponential distribution $P[s]$ above, write a decoder (on paper, not programming) that for each value of x returns the most likely value of s as s_{est} . How does it differ from the linear decoder (5.2.12)? Continue to assume Gaussian noise and just return the most likely value. Do not carry out analytical minimization of χ^2 .*

There is extended discussion of fly H1 experiments in [179], and a key point is made in Figure 2.2. The conditional mean of the neural output given the stimulus (encoding) is shown in i), and it is clearly nonlinear. The conditional mean of the stimulus given the neural output is shown in h), and it is approximately linear. In this system, nonlinear encoding *does not* imply nonlinear decoding. Exercise 34 illustrates that the opposite can occur: linear encoding can accompany nonlinear decoding. Thus, given linearity of one, you cannot assume it of the other.

Finally we address causality. In encoding, this is simple: spikes are only caused by stimulus events that precede them: a spike at time t_1 is related to $s(t)$ only for $t < t_1$. Similarly, in decoding at time t_0 , we may only use spikes that occurred for $t_i < t_0$, but each of them was caused by the stimulus for $t < t_i$. What does this imply about the stimulus at t_0 if the most recent spike preceded it by $t_0 - t_n$?

There are two solutions to this problem. The first is the most intuitively obvious: decode with a delay. We cannot determine the stimulus at the current time t_0 , but if we look back over duration τ_0 , we can estimate the stimulus if there is at least one spike in $[t_0 - \tau_0, t_0]$, and the more spikes in this interval, the better we can do. This suggests making τ_0 large, to get more accurate estimates. But there is usually a competing pressure for speed. A fly with $\tau_0 = 200$ ms could not turn quickly, and most experiments indicate that $\tau_0 \approx 30$ ms for H1 neuron decoding. (We shall meet another *speed-accuracy tradeoff* in human decision making in §6.) The second solution is to assume that the stimulus has correlations that extend forward in time: if you know the stimulus at time t_1 , then you know *something* about it for $t > t_1$ (a sort of continuity). With correlation time τ_c , the best guess for the stimulus at time $t > t_1$ is $s(t_1)$ decaying to the average value s_{av} of the stimulus with time constant τ_c . (Note that $s_{av} = 0$ if s is a zero mean Gaussian as assumed above). In this way, all spikes that occur before t_0 can be used to estimate $s(t_0)$. Because of causality, we must use either delays in decoding, correlations in the stimulus, or both.

We can now create a simple model for decoding. We estimate the stimulus via a linear response function, which takes both encoding by a given neuron and the stimulus distribution into account. As in linear vibration theory, where we convolve a driving force with an impulse response function to get a particle's trajectory, here we convolve a sequence $\rho(t)$ of delta function spikes with the linear response function to get the stimulus estimate:

$$s_{est}(t) = \int_0^T K_1(\tau) \sum_{i=1}^N \delta(t - \tau - t_i) d\tau = \sum_{i=1}^N K_1(t - t_i). \quad (5.2.13)$$

If estimation is delayed, then we effectively find $s_{est}(t - \tau_0)$.

Higher order terms may be added as in the Wiener filter, but in many systems, including fly H1, this does not help. This is because the spike triggered average $C(\tau)$ provides the best stimulus estimate if there is only a single spike in the observation window: with estimation delay τ_0 , our best filter is

$$K_1(\tau) = C(\tau_0 - \tau), \quad (5.2.14)$$

i.e., $C(\tau)$ run backwards and shifted by τ_0 . The filter K_1 has limited temporal width, related to the stimulus correlation time τ_c . If the product $\langle r \rangle \tau_c$ is small, implying less than one spike on average in each period of length τ_c , the reconstruction at any given time will depend on a single spike. As spike density increases, higher order corrections may be needed, but for many of the systems noted earlier, the reconstruction $s(t)$ given by the STA and the spike in the interval $[t - \tau_0, t]$ is adequate.

If we include stimulus correlations, then spikes before $t - \tau_0$ also affect the reconstruction, since the stimulus estimate at those spike times decays as τ_c to the average stimulus value. If neither delay nor stimulus correlations are included, then the best guess for the stimulus at any time is simply the stimulus average and $K_1(\tau) = 0$. For sufficiently large delays, setting $K_1(\tau) = C(\tau_0 - \tau)$ to zero for $\tau < 0$ has little effect, since $C(\tau) \rightarrow 0$ for large τ : spikes are not correlated to stimulus values far enough in the past. The part of $K(\tau)$ set to 0 is already very close to 0 for τ_0 large enough for many systems, and $K(\tau)$ is set to zero for $\tau < 0$ because of causality: spikes occurring after time t cannot affect our estimate at $t - \tau_0$. The previous reasoning merely states that this imposition of causality does not affect accuracy much. The construction of $K(\tau)$ from $C(\tau)$ is shown in Figure 79.

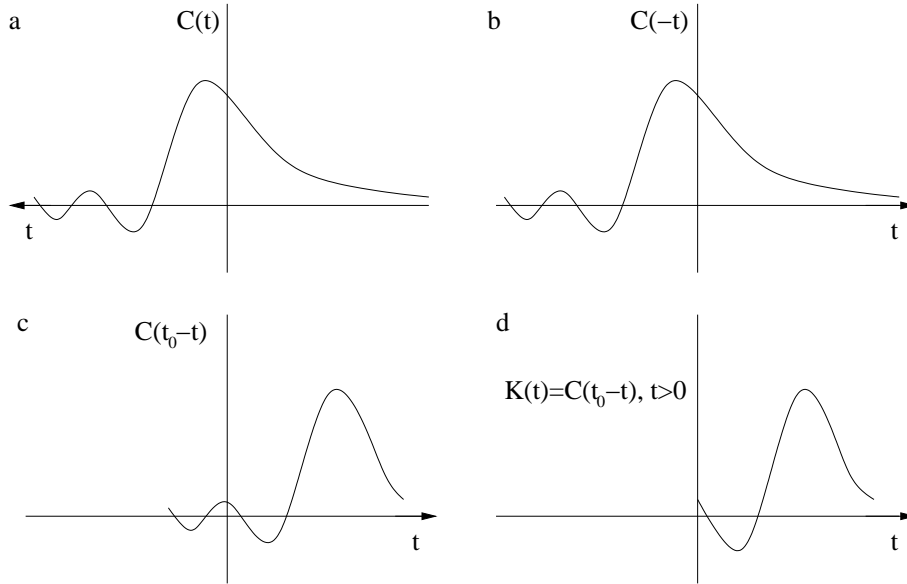


Figure 79: a) The spike triggered average $C(\tau)$. b) Flipping time to get $C(-\tau)$. c) Shifting $C(-\tau)$ by τ_0 to introduce the estimation delay. d) Setting $K(\tau) = 0$ for $t < 0$ to enforce causality in estimation.

A final interpretation of the linear response function K_1 is that it serves as a low pass filter. High frequency components of the spike train are attenuated, so the reconstruction is insensitive to small changes in the spike timing. The fly H1 filter attenuates frequencies above approximately 25 Hz. Provided the relevant information is carried below the cutoff frequency, this also filters out high frequency noise in the neural system, a desirable feature which helps linear decoding work. Thus, in many cases linear models that do not work for encoding do work remarkably well for decoding. For more information see [179, §2.3] and [50, §3.4].

Another model for decoding is explored in [179, §2.2]. Briefly, one computes the stimulus given a spike pattern. For instance, what is the average stimulus, given a two spikes separated by 20 ms

with none in between? During extended stimulus presentations, patterns of two spikes, or a single spike preceded by 25 ms of silence, etc., appear many times. Random sampling of the stimuli that correspond to these responses gives a response-conditional ensemble for each response. This can be done in depth, looking at many different responses, which allows a type of decoding that finds the appropriate stimulus given a response. An examination and explanation of literature for this approach, some of which is referenced in [179, §2.2], would make a good semester project.

5.3 Information theoretic approaches

Information theory gives us a principled way to measure information transmission in *bits/second* or *bits/spike*, in much the same way that an internet connection might have had a data rate of 56 *kbits/sec*, back when the first draft of these notes began. This allows us to ask interesting questions which would be hard to formulate without a proper way of quantifying information: Can we quantify in bits the information that a spike train conveys about the environment? Is there a maximum possible information transmission rate for a neuron, and how close are real neurons to being optimal? Is there redundancy between the information transmitted by different neurons? How much does noise corrupt the information transmitted by a neuron, and can this lost information be recovered by looking at more than one neuron? Do groups of neurons transmit information optimally?

Information theory also gives ways of evaluating how well a given encoding/decoding model captures recorded data. One can compare the information transmission rate of a model with a model-independent estimate of how much information is actually being transferred. For example, one can ask how much information is lost by characterising a spike train only by its firing rate $r(t)$, and whether a model which tries to capture certain structures in the spike train (such as bursts) serves to transmit information more efficiently. The mathematical theory of information began with Shannon's classic paper [195] of 1948, in which entropy was introduced to describe signal transmission over noisy channels, cf. [196]. (Entropy arose in statistical physics in the 19th century, and had already been characterized by Boltzmann as missing information.) In the same year Wiener proposed entropy as a useful descriptor for sensory receptors and information transmission in the nervous system [229, especially Chap. III]. Also see [179, Chap. 3] for more background on information theory and further examples of its uses in neuroscience.

5.3.1 Entropy and missing information

The most common information-theoretic quantity is the *entropy* $S(p)$ of a *probability distribution* p . Specifically, given a discrete random variable X that can take any one of N distinct values X_j with probabilities $p_j = p(X_j)$ (e.g., X_j = letters in a message, or words in a dictionary), the entropy of p is defined as

$$S(p) = -k \sum_{j=1}^N p_j \log p_j, \quad (5.3.1)$$

where k is a constant which may be eliminated by choosing the base of the logarithm. The convention is to define *bits* of information by taking logs to the base 2:

$$S(p) = - \sum_{j=1}^N p_j \log_2 p_j \quad \text{or, a little more generally, } S(p) = - \sum_x p(x) \log_2 p(x). \quad (5.3.2)$$

In this context, entropy measures the average number of yes/no questions needed to determine which element X_j has been chosen in drawing a sample, assuming that for each trial an oracle chooses one particular element (with probability p) for us to guess. Intuitively, it is a measure of (the log of) the effective number of elements, taking into account that p may put greater weight on some elements than others.

If the distribution p is uniform, its entropy is simply \log_2 of the number of elements; thus, for a fair coin $S = -2 \times (1/2) \log_2(1/2) = \log_2 2 = 1$. One bit of information is revealed when the coin lands: heads, or tails. For a fair six-sided die, we have

$$S = - \sum_{j=1}^6 \frac{1}{6} \log_2 \left[\frac{1}{6} \right] = \log_2 6 = \frac{\ln 6}{\ln 2} = 2.58496 \dots \quad (5.3.3)$$

(In the final step we recall the definition of logarithms to the base a , the natural logarithm $\log_e x = \ln x$, and the formula $\log_a x = \ln x / \ln a$, which we shall use again later.) For a fixed number of elements, non-uniformity in the distribution p reduces entropy: e.g., for a loaded die in which two numbers come up with probabilities each $1/4$ and the remaining four with probabilities $1/8$, we have

$$S = - \left\{ 2 \times \frac{1}{4} \log_2 \left[\frac{1}{4} \right] + 4 \times \frac{1}{8} \log_2 \left[\frac{1}{8} \right] \right\} = \frac{\log_2 4 + \log_2 8}{2} = 2.5. \quad (5.3.4)$$

Non-uniform probabilities effectively reduce the number of alternatives that are likely to appear.

The definition (5.3.2) respects the notion of independence introduced at the end of §5.1.2, $p(x, y) = p(x)p(y)$, in that for any distribution of two independent variables

$$S[p(x, y)] = S[p(x)] + S[p(y)] : \quad (5.3.5)$$

thus, entropies are *additive* (iteration of (5.3.5) leads to an analogous expression for N variables). However, one might ask if other definitions of entropy would suffice. In fact, the additivity constraint (5.3.5) defines $S(p)$ uniquely if we also require that (1): $S(p)$ be a continuous function of the probabilities and (2): the entropy σ_k of the uniform distribution over k outcomes increases with k . The following exercises help you prove that under these assumptions, $S(p)$ must be given by Eqn. (5.3.2).

Exercise 35. Use the expression $S[p(x, y)] = S[p(x)] + \sum_x p(x) S[p(y|x)]$ to show that the entropies of independent variables add, and in particular that $\sigma_{k^n} = n\sigma_k$. (Recall that $p(y|x)$ denotes the probability of y , given x .)

Exercise 36. It is a fact that for an arbitrarily large integer n , there exists an integer $m(n)$ such that $2^{m(n)} \leq k^n < 2^{m(n)+1}$. Use this and the assumption that σ_k increases with k to show that $\sigma_k = \log_2 k$.

Exercise 37. Using $S[p(x, y)] = S[p(x)] + \sum_x p(x) S[p(y|x)]$, show that the entropy of a probability distribution in which $p_i = k_i / \sum_i k_i$ (i.e. p_i is a rational number for all i) is given by (5.3.2). Can you now prove that (5.3.2) holds for all discrete probability distributions?

Continuous distributions can be accommodated by replacing the sum with an integral to define *differential entropy* $S(p) = - \int p(x) \log p(x) dx$, where $p(x)$ is a probability density with respect to some measure dx . For the Gaussian distribution of Eqn. (5.1.11), this gives

$$\begin{aligned} S[N(\mu, \sigma)] &= - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \log_2 \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \right\} dx \\ &= \dots = \frac{1}{2} \log_2 (2\pi e \sigma^2) \text{ bits,} \end{aligned} \quad (5.3.6)$$

where \dots represents a calculation that is detailed in [179, Appendix A.9]. This example emphasizes the fact that entropy measures variability: up to an additive constant, $S[N(\mu, \sigma)] = \log_2 \sigma$. Also note that the entropy is independent of the mean μ .

In general, entropy is well-defined only for discrete probability distributions over finite numbers of elements. In particular, differential entropy is not invariant under changes of variables such as changing the units of x . This is troublesome, since entropy should be a function of the probability distribution alone, and not of the underlying element space. As noted in [179, §3.1], the entropy of a voltage distribution should not change by a factor of $3 \log_2(10)$ if one goes from mV to V in measuring variances! Moreover, differential entropies can be negative. The underlying problem is that the number of states available to a continuous variable x is infinite, even when the support of $p(x)$ is finite ($p(x) \equiv 0$ for $x \notin [a, b] \subset \mathbb{R}$).

Fortunately, *differences* of differential entropies *are* invariant under changes of variables. For example, comparing Gaussians with variances σ_1 and σ_2 the difference is

$$S[p_1(x)] - S[p_2(x)] = \frac{\log_2 (2\pi e \sigma_1^2) - \log_2 (2\pi e \sigma_2^2)}{2} = \log_2 \left(\frac{\sigma_1}{\sigma_2} \right); \quad (5.3.7)$$

all the constants vanish, as would multiplicative factors due to changes in units. This also reveals that doubling the standard deviation increases entropy by 1 bit, a consequence of taking logs to the base 2. Finally, note that mutual information, introduced in §5.3.4, can be expressed as differences of entropies.

5.3.2 Entropy of spike trains and spike counts

We now illustrate an application to neuroscience by computing the entropy of a spike train, following [179, §3.1.2]. Suppose that we have a record of length T from a neuron firing at rate r , or, more precisely, with spike count rate r , according to Eqn. (5.1.4). We further assume that the record is discretized with precision (bin size) Δt , small enough so that at most one spike falls in each bin (cf. §5.1.4). Labeling bins with and without spikes by 1's and 0's respectively, a spike train becomes a binary word of length $N = T/\Delta t$ containing rT 1's, and the number of distinct possible words is given by the binomial coefficient (N choose rT):

$$\binom{T/\Delta t}{rT} = \frac{(T/\Delta t)!}{(rT)!(T/\Delta t - rT)!} = N_{tot}. \quad (5.3.8)$$

If each spike train is equally probable, the probability of any given word $i \in [1, N_{tot}]$ is $p_i = 1/N_{tot}$, from which we may compute the entropy of the spike train

$$\begin{aligned} S_{train} &= - \sum_{i=1}^{N_{tot}} p_i \log_2(p_i) = - \sum_{i=1}^{N_{tot}} \frac{1}{N_{tot}} \log_2 \left(\frac{1}{N_{tot}} \right) = \log_2 \left[\frac{(T/\Delta t)!}{(rT)!(T/\Delta t - rT)!} \right] \\ &= \frac{\ln[(T/\Delta t)!] - \ln[(rT)!] - \ln[(T/\Delta t - rT)!]}{\ln 2}. \end{aligned} \quad (5.3.9)$$

If $N = T/\Delta t, rT$ and $N - rT$ are all large (e.g., as $T \rightarrow \infty$), we may simplify (5.3.9) by using Stirling's approximation, which holds for large N :

$$N! \sim \sqrt{2\pi} N^{N+1/2} e^{-N}, \text{ or } \ln(N!) = N[\ln(N) - 1] + \mathcal{O}(\ln(N)). \quad (5.3.10)$$

Substituting (5.3.10) into the three factorials of (5.3.9), neglecting the $\mathcal{O}(\ln(\cdot))$ terms, and further simplification detailed in [179, §3.1.2] yields

$$\begin{aligned} S_{train} &\approx - \frac{T}{\Delta t \ln 2} [r\Delta t \ln(r\Delta t) + (1 - r\Delta t) \ln(1 - r\Delta t)] \\ &= - \frac{T}{\Delta t} [r\Delta t \log_2(r\Delta t) + (1 - r\Delta t) \log_2(1 - r\Delta t)]. \end{aligned} \quad (5.3.11)$$

The assumption of ≤ 1 spike/bin implies that $r\Delta t < 1$, and so $S_{train} > 0$.

Note that $S \propto T$, where T is the duration of the spike train, and if the probability $r\Delta t$ of observing a spike in any bin is small, we may use the approximation $\ln(1 - r\Delta t) = -r\Delta t + \mathcal{O}((r\Delta t)^2)$ to estimate the *entropy rate* of the spike train:

$$\frac{S_{train}}{T} \approx \frac{r\Delta t[1 - \ln(r\Delta t) + \mathcal{O}(r\Delta t)]}{\Delta t \ln 2} \approx \frac{r[\ln(e) - \ln(r\Delta t)]}{\ln 2} = r \log_2 \left(\frac{e}{r\Delta t} \right) \text{ bits/sec}. \quad (5.3.12)$$

Dividing (5.3.12) by the mean firing rate r , we obtain the *entropy per spike* in units of bits/spike. Fig. 80(a) plots this quantity as a function of the timing precision Δt . Note that the entropy can exceed 1 bit/spike, especially for small Δt . The information per bin cannot exceed 1 bit, but empty bins also provide information. More precisely, interspike intervals (ISIs) are distributed with mean $1/r$ sec and each ISI is measured with accuracy $\sim \Delta t$, and therefore chosen from $\sim 1/r\Delta t$ possibilities. Each ISI, and hence each spike, is thus associated with $\sim \log_2(1/r\Delta t)$ bits of entropy.

Spike train entropy provides a standard against which to gauge the performance of different coding schemes, including real neural codes. For example, we can compare the expressions (5.3.11) and (5.3.12), which assume that spike times are tracked with precision Δt , with a scheme in which we simply count the total number of spikes in a (possibly large) window T , or equivalently, measure the average firing rate as in Eqn. (5.1.4). The former case corresponds to a *spike timing code*, and the latter to a *spike rate code* or simply *rate code*, in which information is carried by the *average spike count* $n = rT$ or its associated rate r . Letting $P_T[n]$ denote the probability of finding n spikes in a window of length T as in §5.1.4, we require the entropy of the *spike count distribution*:

$$S_{count} = - \sum_n P_T[n] \log_2(P_T[n]). \quad (5.3.13)$$

Since $P_T[n]$ is a probability distribution with mean rT , it must satisfy

$$\sum_n P_T[n] = 1 \quad \text{and} \quad \sum_n n P_T[n] = rT, \quad (5.3.14)$$

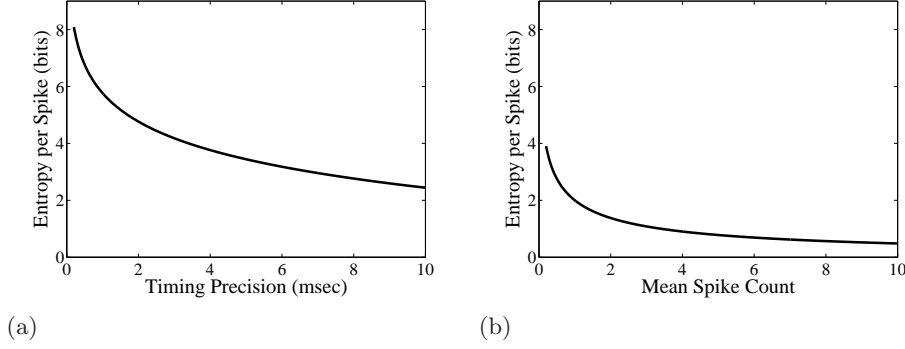


Figure 80: Entropy of a spike train. (a) Entropy per spike vs. bin width Δt (Eqn. (5.3.12) divided by r) for $r = 50$ spikes/sec. (b) Entropy per spike vs mean spike count rT (Eqn. (5.3.21) divided by rT). Figure replotted from [179].

and although these two constraints do not uniquely specify $P_T[n]$, we can use them to find the distribution that maximizes the entropy. This is the most random description of spike count probability consistent with our prior knowledge, and it will provide an upper bound for the entropy of spike counts from any real neuron.

To maximize (5.3.13) under the constraints (5.3.14), we seek a critical point of the functional

$$G(P_T[n]) = - \sum_n P_T[n] \log_2(P_T[n]) - \lambda_1 \left(\sum_n P_T[n] - 1 \right) - \lambda_2 \left(\sum_n n P_T[n] - rT \right), \quad (5.3.15)$$

where λ_1 and λ_2 are Lagrange multipliers. This is a problem in the calculus of variations, introductions to which can be found in [228, 83]. Much as in multivariable calculus, to locate a maximum we take the first and second derivatives to find analogs of the gradient vector and Hessian matrices, set the former equal to zero to locate a critical point, and check that the latter is negative definite to verify that this is a (local) maximum. Specifically, to find the critical point (here, a “critical function”) of (5.3.15), we perturb the function $P_T[n]$ by adding a small variation $\epsilon h[n]$, differentiate $G(P_T[n] + \epsilon h[n])$ w.r.t. ϵ (cf. Eqn. (5.3.15)), and set $\epsilon = 0$. A long, explicit calculation is displayed in [179, Appendix 11], in which the second variation is also computed. The expression for the first variation is found to be:

$$\delta G = - \sum_n \left\{ \frac{1 + \ln(P_T[n])}{\ln 2} + \lambda_1 + \lambda_2 n \right\} h[n]. \quad (5.3.16)$$

Since $h[n]$ is arbitrary, we may choose it to be non-zero on any integer n and zero on all others, implying that the expression within the curly brackets must vanish for all values of n , if $P_T[n]$ is to be a critical point. After rearrangement, this yields

$$P_T[n] = \frac{1}{Z} e^{-\lambda n}, \quad (5.3.17)$$

in which we have subsumed the Lagrange multipliers λ_1 and λ_2 into the constants Z and λ . We have determined that an exponential distribution maximizes the spike count entropy.

So far, we have not specified the limits of the sums on n in Eqns. (5.3.13-5.3.16). It is possible (no matter how unlikely) that no spikes fall in the interval $[0, T]$, so the lower limit is $n = 0$. However,

we cannot assign an upper bound to n , for although neural firing rates are bounded above, more and more spikes may be observed as T grows. Moreover, in continuing to solve for Z and λ it is simpler to allow $n \rightarrow \infty$.

To find the normalization constant Z , we use the first of the constraints (5.3.14), $\sum_{n=0}^{\infty} e^{-\lambda n}/Z = 1$, and sum the geometric series to obtain

$$Z = \sum_{n=0}^{\infty} (e^{-\lambda})^n = \frac{1}{1 - e^{-\lambda}}. \quad (5.3.18)$$

The second constraint can then be written

$$\frac{1}{Z} \sum_{n=0}^{\infty} n e^{-\lambda n} = rT. \quad (5.3.19)$$

To sum this series, it is helpful to note that $\frac{\partial}{\partial \lambda} e^{-\lambda n} = -n e^{-\lambda n}$, so that we can rewrite (5.3.19) as follows:

$$rT = \frac{1}{Z} \sum_{n=0}^{\infty} (-1) \frac{\partial}{\partial \lambda} e^{-\lambda n} = -\frac{1}{Z} \frac{\partial}{\partial \lambda} \sum_{n=0}^{\infty} (e^{-\lambda})^n = -\frac{1}{Z} \frac{\partial}{\partial \lambda} \left(\frac{1}{1 - e^{-\lambda}} \right) = \frac{e^{-\lambda}}{1 - e^{-\lambda}}, \quad (5.3.20)$$

where (5.3.18) was used in the last equality. Thus, $\lambda = \ln(1 + 1/rT)$ (> 0 , as required), and via (5.3.18) we find that $Z = 1 + rT$. At last, substituting $P_T[n]$ into (5.3.13) and using almost everything found above, we obtain

$$\begin{aligned} S_{count}^{max} &= - \sum_{n=0}^{\infty} P_T[n] \log_2(P_T[n]) = \dots = \frac{1}{\ln 2} \left(\ln Z \sum_{n=0}^{\infty} \frac{e^{-\lambda n}}{Z} + \lambda \sum_{n=0}^{\infty} \frac{n e^{-\lambda n}}{Z} \right) \\ &= \frac{\ln Z + \lambda rT}{\ln 2} = \log_2(1 + rT) + rT \log_2 \left(1 + \frac{1}{rT} \right) \text{ bits}. \end{aligned} \quad (5.3.21)$$

Dividing S_{count}^{max} by the mean spike count rT , in Fig. 80(b) we plot the *maximum entropy per spike* as a function of rT . Note that it decreases with T , implying that the spike train's capacity to carry information declines as the time resolution coarsens. Indeed, as $rT \rightarrow \infty$, the available information capacity per spike $S_{count}^{max}/rT \rightarrow 0$. This is as expected, for firing rates must vary with time if they are to carry much information, and we are averaging over a window of length T . Indeed, if we count spikes in windows with lengths equal to the mean ISI, so that $rT = 1$, then the entropy in S_{count}^{max} is precisely 2 bits/spike: rate codes can carry more than 1 bit per/spike if spike rates are measured on time scales comparable to ISIs. More significantly, Eqns. (5.3.11) and (5.3.21) reveal that the entropy of the spike train is typically greater than the entropy of the spike count, because keeping track of individual spike timing allows for a larger set of available states, and thus more available information.

Exercise 38. Quantify the claim made immediately above by computing the behaviors of S_{train} and S_{count}^{max} as T increases for fixed r and Δt . In particular, how does S_{count}^{max} depend on T as $T \rightarrow \infty$?

Finally, if we count the numbers of spikes in very small time windows so that $rT \ll 1$, the maximum entropy rate of the spike count becomes

$$\frac{S_{count}^{max}}{T} = \frac{\ln(1 + rT) + rT \ln(1 + 1/rT)}{T \ln 2} = \frac{r[1 - \ln(rT) + \mathcal{O}(rT)]}{\ln 2} \approx r \log_2 \left(\frac{e}{rT} \right) \text{ bits/sec} : \quad (5.3.22)$$

this is just the entropy rate of the full spike train (5.3.12), with the bin size Δt replaced by T . Counting spikes in small enough windows blurs the distinction between rate and timing codes.

What rate or timing codes actually tell us about sensory inputs depends on the time scales of the stimuli, and on the range of firing rates available to the relevant neurons. If a signal (in the natural world or laboratory) varies sufficiently slowly, then it makes sense to divide the spike train into windows that each contain many spikes. The resulting time-varying spike rate can presumably report quasi-static descriptors of the signal averaged over each window. In contrast, a timing code can probe brief transients and sudden events. In this case, the distinction between rate and timing codes is clear. However, if stimuli vary on time scales comparable to ISIs, then the natural time windows contain few spikes, and the rate code transforms smoothly into a timing code, as shown in the calculations above.

In §6 we shall describe models of evidence accumulation over periods of $\mathcal{O}(1)$ sec that rely on firing rates. In contrast, sound localization via binaural sensing requires precision on the order of 10 – 200 microseconds: 1 – 2 orders of magnitude shorter than an action potential. This strongly suggests that a timing code is used to transmit interaural time differences (ITDs) between the two ears. The classical theory of Jeffress [118] posits the use of axonal delay lines from sensory cells in the two ears. Depending upon the angle of the sound source relative to the head, spikes traveling along the delay lines coincide at different locations on an array of coincidence detectors. These convert ITDs into spike rates, transforming the timing code to a place map in which firing rates of cells signal the probable direction of the source, much as in the orientation-selective neurons in visual cortex [114] (cf. [231, §§6.2, 7.5]. Neural circuits consistent with Jeffress' theory have been found in animals such as the barn owl and chicken, but other mechanisms have also been identified in birds, mammals and reptiles, particularly regarding details of the rate coding of ITDs. Acoustic coupling via the mouth cavity can increase ITDs. A distinct system using interaural level differences (ILDs) also exists, and yet another principle involving monaural frequency analysis is used for localization in the vertical plane.

Recent reviews of the sound-localization literature appear in [87, 9]. In this field details of ion channels, dendrites and axons in single cells, timescales of synapses, balances of excitation and inhibition, and circuit architectures interact with the mechanics of sound waves to form a rich environment both for nature's evolutionary solutions, and for theorists' and modelers' contributions. Some examples of the latter can be found in [3, 208, 119], and in particular [95] develops an information-theoretic analysis of optimal coding strategies to explain different mechanisms on the basis of head size and frequency ranges. The whole area would be a good source for final projects.

Exercise 39. *Redo the computations leading to $P_T[n]$ above without the second constraint of (5.3.14), i.e., find the distribution that maximizes $S(p[n]) = -\sum_n p[n] \log_2 p[n]$ subject only to the normalization condition $\sum_n p[n] = 1$. What additional assumptions on the sums, if any, are necessary for your calculations to make sense? Does the distribution that you have found seem reasonable for describing neural spike counts?*

5.3.3 Relative entropy

Relative entropies (also known as the Kullback-Leibler divergence) are defined in a rather broad context. They can be concisely developed from first principles. Suppose we are playing a game of ‘20 questions’ on the interval $[0, 1]$: there is a particular number a we need to find by asking an oracle yes/no questions such as: ‘Is $a < 1/3$?’. We start off not knowing anything about a , so we assume a to have uniform probability r over $[0, 1]$. Suppose further that the actual value of a is drawn from a probability distribution p . For example if p is a Dirac delta function $\delta(0)$ at $a = 0$, the oracle will answer our question by saying ‘ $a < 1/3$ with probability 1’. Let us try to quantify the information gained by asking a set of such questions. We want the relative entropy $\mathcal{I}(p||r)$ ¹² to be solely a function of our prior expectations $r(\text{answers})$ and of the probabilities $p(\text{answers})$ implied by the oracle’s responses. If these are very different from what was expected, then the question was informative.

The information gain should have some basic properties. If the oracle always answers what we expected a priori, then we already know what the oracle knows ($p = r$) and the information gain should be $\mathcal{I}(r||r) = 0$. We would also like the order in which we ask questions not to change how much information is gained. We can start by asking a certain question X , thereby gaining information $\mathcal{I}(p_X||r_X)$. If the answer is $X = x$ and then we proceed to ask another question Y , the quantity of *new* information gained will be $\mathcal{I}(p_{Y|X=x}||r_{Y|X=x})$. Since each answer to X occurred with probability p_X , the *average* information gain of Y is $\sum_X p_X \mathcal{I}(p_{Y|X}||r_{Y|X})$. To ensure that the order in which questions are asked does not matter, we require:

$$\mathcal{I}(p_{X,Y}||r_{X,Y}) = \mathcal{I}(p_X||r_X) + \sum_X p_X \mathcal{I}(p_{Y|X}||r_{Y|X}) = \mathcal{I}(p_Y||r_Y) + \sum_Y p_Y \mathcal{I}(p_{X|Y}||r_{X|Y}). \quad (5.3.23)$$

Here $p_{X,Y}$, $p_{X|Y}$, $p_{Y|X}$, etc. denote the joint and conditional probabilities, as defined in §5.1.2. We do not distinguish between the discrete and continuous cases ($P[X, Y]$ and $p(x, y)$, etc.).

Surprisingly, there is only one continuous function of p and r which satisfies $\mathcal{I}(r||r) = 0$ and (5.3.23), up to a multiplicative constant. The multiplicative constant is usually chosen so that one *bit* is the information gained by asking a yes/no question for which our prior expectation is $1/2$ ‘yes’ $1/2$ ‘no’ and the actual answer is ‘yes with probability 1’, (this is the most informative yes/no question possible). The following exercises help you prove that under these assumptions, the relative entropy must have the expression:

$$\mathcal{I}(p_X||r_X) = \sum_X p_X \log_2 \left(\frac{p_X}{r_X} \right) \text{ bits}. \quad (5.3.24)$$

Exercise 40. Suppose we are trying to guess a number $a \in [0, 1]$, and our prior expectations over $[0, 1]$ are uniform. Use $\mathcal{I}(r||r) = 0$ and (5.3.23) to show that if the answer to the question ‘Is $a < t^2$?’ is ‘Yes with probability 1’ for some $0 < t < 1$, then $\mathcal{I}(a < t^2) = 2 \mathcal{I}(a < t)$.

Exercise 41. Show that if the answer to the question ‘Is $a < 1/2$?’ is ‘Yes with probability 1’, then this question contains one bit of information. Now show that if the answer to ‘Is $a < t$?’ is ‘Yes with probability 1’ for some $0 < t \leq 1$, then $\mathcal{I}(a < t) = \log_2 \frac{1}{t}$. [Hint: construct a decreasing sequence t_n converging towards t for which $\mathcal{I}(a < t_n)$ is known for all n .]

¹²the double bar $||$ notation is conventional.

Exercise 42. Suppose the answer to the question X ‘Is $a < t$?’ is ‘Yes with probability α ’. By expanding the yes/no question game to guessing $(a, b) \in [0, 1]^2$ for a well-chosen probability distribution $p(a, b)$ and a uniform prior, show that $\mathcal{I}(p_X || r_X)$ satisfies (5.3.24). [Hint: use (5.3.23) once more.] Can you generalize to show that (5.3.24) holds for questions with more than 2 answers or for more than one question?

One can show that (5.3.24) continues to define the gain in information for very general p and r (as long as p is absolutely continuous with respect to (w.r.t.) r) by replacing sums \sum with integrals \int or expectations \mathbb{E} . The questions can be much more general than yes/no questions. In fact, it is not necessary to define questions at all: $\mathcal{I}(p || r)$ defines the information someone gains by being told that the true probability distribution is p when he thought it was r .

Exercise 43. Using Jensen’s inequality¹³, show that $\mathcal{I}(p || r)$ is always ≥ 0 , and only $= 0$ if $p = r$. Use this to show that asking more questions always increases the information gain.

5.3.4 Mutual information

We can now define the *mutual information* $\mathcal{MI}(X, Y)$ which measures the information shared between two random variables, and can be written in a number of equivalent ways:

$$\begin{aligned}\mathcal{MI}(X, Y) &= \mathcal{I}(p_{X,Y} || p_X p_Y) = \int \int p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right] dx dy \\ &= \int p(y) \mathcal{I}(p_{X|Y} || p_X) dy = \int p(y) \int p(x|y) \log_2 \left[\frac{p(x|y)}{p(x)} \right] dx dy \\ &= \int p(x) \mathcal{I}(p_{Y|X} || p_Y) dx = \int p(x) \int p(y|x) \log_2 \left[\frac{p(y|x)}{p(y)} \right] dy dx.\end{aligned}\tag{5.3.25}$$

The expressions in (5.3.25) can be derived by use of the conditional entropy

$$S(x|y) = - \int p(x|y) \log_2 p(x|y) dx,\tag{5.3.26}$$

which quantifies the information available, or remaining variability, in x given the observation y . The gain in information due to observing y , and averaging over many such observations, is therefore

$$\mathcal{MI}(X, Y) = \int [S(x) - S(x|y)] dy.\tag{5.3.27}$$

Exercise 44. Derive the three integral expressions in (5.3.25) from Eqns. (5.3.26-5.3.27) and the relations among the joint and conditional probabilities for x and y .

For example, we can calculate the mutual information between the spiking of a neuron and a stimulus $s(t)$ that leads up to the spike, as we expect spikes to carry information about the stimulus

¹³Jensen’s inequality may be stated as follows. For a real-valued, convex function ϕ , numbers x_1, x_2, \dots, x_n and weights $a_j > 0$, $\phi \left(\frac{\sum_j a_j x_j}{\sum_j a_j} \right) \leq \frac{\sum_j \phi(a_j)}{\sum_j a_j}$.

through $p(s(t < t_0) | \text{given a spike at } t_0) \stackrel{\text{def}}{=} p_{s|spike}$. This calculation can be done in two different ways. If r is the average firing rate of the neuron, then the probability of having a spike in an interval of length Δt is $\langle r \rangle \Delta t$ and we can write:

$$\mathcal{MI}(s, spike \in [t; t + \Delta t]) = \langle r \rangle \Delta t \mathcal{I}(p_{s|spike} || p_s) + (1 - \langle r \rangle \Delta t) \mathcal{I}(p_{s|no\ spike} || p_s). \quad (5.3.28)$$

Note that $p_{s|no\ spike} = p_s$ = the probability distribution of the stimulus. This is because in our idealization spikes are mathematical points (they take no time to happen), so conditioning on having no spike is equivalent to not conditioning at all. (In fact the argument that shows this is subtle, since one must take the limit $\Delta t \rightarrow 0$.) This implies that the second term of (5.3.28) is zero. Dividing through by Δt gives the mutual information between the stimulus and a spike in bits/second, and dividing by $\langle r \rangle$, we obtain the mutual information as

$$\mathcal{MI}(s, spike) = \mathcal{I}(p_{s|spike} || p_s) \text{ bits/spike}. \quad (5.3.29)$$

The expression (5.3.29) makes a lot of sense, since it is saying that the information gained by knowing there was a spike is determined by how different the distribution of stimuli is close to a spike compared to the overall stimulus distribution. In practice, evaluating this quantity requires an estimate of $p_{s|spike}$ and p_s , which may or may not be feasible, depending on the stimulus. We can do the same calculation by first integrating over stimuli. This yields

$$\mathcal{MI}(s, spike) = \int p(s) \mathcal{I}(p_{spike|s} || p_{spike}) ds \text{ bits/second}. \quad (5.3.30)$$

Once again, measuring $p_{spike|s}$ could be difficult for high-dimensional stimuli, but as usual one can present the same stimulus many times while recording a neuron's response in order to get an estimate of $p_{spike|s(t)} = r(s(t))$. Recalling the definition of ensemble averages in Eqns. (5.1.5-5.1.6) of §5.1.1, we replace $p_{spike|s(t)}/p_{spike}$ with $r(t)/\langle r \rangle$ and divide through by $\langle r \rangle$ to obtain:

$$\mathcal{MI}(s, spike) = \frac{1}{T} \int_0^T \left[\frac{r(t)}{\langle r \rangle} \log_2 \frac{r(t)}{\langle r \rangle} \right] dt \text{ bits/spike}. \quad (5.3.31)$$

For early visual neurons such as retinal ganglion cells or the H1 neuron in the fly, the amount of information that each spike carries about the visual scene is in the range of 1 – 6 bits/spike. Note that this calculation does not take into account correlations between spikes. In practice, spikes that occur in close succession often convey overlapping pieces of information about the stimulus, and the overall information transferred by the spike train is consequently less than the sum of informations transferred by each spike separately. This need not always be the case: it is possible for the total information that several spikes carry about the stimulus to be more than the sum of each separate spike's information. For example, In the spike train code of §5.3.2 information per bin cannot exceed 1 bit, but entropies can exceed 1 bit/spike, because interspike intervals are also monitored.

As noted in §5.3.1, mutual information can be expressed as differences of entropies:

$$\begin{aligned} \mathcal{MI}(X, Y) &= \int \int p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right] dx dy = S(p_X) + S(p_Y) - S(p_{X,Y}) \\ &= \int p(y) \int p(x|y) \log_2 \left[\frac{p(x|y)}{p(x)} \right] dx dy = S(p_X) - \int p(y) S(p_{X|Y}) dy \\ &= \int p(x) \int p(y|x) \log_2 \left[\frac{p(y|x)}{p(y)} \right] dy dx = S(p_Y) - \int p(x) S(p_{Y|X}) dx. \end{aligned} \quad (5.3.32)$$

Since entropy differences are well defined for continuous random variables, these expressions may also be used in such cases.

Exercise 45. *For a probability distribution over a finite number of elements, derive an expression for $S(p)$ as a function of $\mathcal{I}(p||u)$, where u is the uniform distribution. What goes wrong when trying to generalize to an infinite number of elements?*

For more examples of these ideas in action in neuroscience, see [21] and [179, Chap. 3].

6 Models of decision making (≈ 2 weeks)

In this final part of the course we continue with phenomenological models, but return to ones inspired by subsets of neural circuitry within the brain. Such *connectionist* or *firing-rate* models have already appeared in Examples 6 and 7 of §2.4, and in §2.5. They are biophysically motivated in the sense that the variables and parameters that define them represent quantities such as firing rates of groups of neurons, synaptic connection strengths, bias currents, etc., but in most cases a satisfactory derivation from mechanistic H-H type models, is lacking or incomplete. However, these “high level” models address phenomena on the scale of brain areas (cf. §1.1 and §3.1), they provide direct links to behavioral observables, and derivations from simplified integrate and fire spiking models are available, as we describe in §6.2.

In this chapter we focus on decision making or signal detection in simple situations in which one of two known stimuli appears, partially buried in noise, and the task is to correctly identify it. A major new aspect, that requires mathematical tools beyond those introduced in §2, is the inclusion of random inputs. After introducing a simple decision task and a model for it in §6.1, and describing an optimal decision strategy in §6.3, we therefore sketch necessary background material on random processes and stochastic differential equations in §6.4. We then return to the optimal procedure in §§6.5-6.6, showing how neural network models approximate it, and discussing behavioral experiments that test human ability for optimal performance.

6.1 Two-alternative forced-choice tasks

Drawing primarily from a recent article [17], we briefly review a canonical behavioral experiment: the two-alternative forced-choice (2AFC) task. Choosing between two alternatives vastly simplifies typical cognitive tasks, but we focus on it for several reasons. First, it represents many problems faced by animals in their natural environments (e.g., whether to approach or avoid a novel stimulus). Pressures for speed and accuracy in such constrained situations may have exerted strong evolutionary influences, thereby optimizing neural decision-making mechanisms. Even if optimality has not been achieved, analyses based on it can provide bounds for possible behaviors. Second, a wealth of human behavioral data has motivated formal modeling of the dynamics and response outcomes in 2AFC tasks (e.g. [136, 172, 36, 174, 203] and see additional references in [17]). Finally, neuroscientists can now directly monitor neuronal dynamics in primates trained to perform the same decision tasks as human subjects, and assess their relationships to task performance. In

many cases, neural and behavioral data are converging to support mathematical models such as those described below (e.g. [93, 186, 194, 78, 173, 203]).

In a common version of the 2AFC task, subjects must identify the direction of a coherently-moving subset of dots embedded in a random motion field [24]. The stimulus appears after a period during which the subject fixates on a central dot, and the response is typically signalled by button pushes for humans, and eye saccades to left or right for monkeys. Key parameters under experimenter control include: i) the stimulus difficulty or signal-to-noise ratio (SNR), which can be manipulated by varying the coherence fraction; ii) whether participants are allowed to respond freely or responses are cued or deadlined; and, as discussed further below, iii) the delay between response and the next stimulus.

Models of 2AFC typically make three fundamental assumptions: i) evidence favoring each alternative is integrated over time; ii) the process is subject to random fluctuations; and iii) the decision is made when sufficient evidence has accumulated favoring one alternative over the other. In the case of fixed duration stimuli, with responses delivered at stimulus offset or signalled by a cue that follows a delay period, the accumulation period is determined by the experimenter rather than the subject. We shall shortly describe a leaky accumulator model [214] that formalizes these ideas. To motivate it, we next sketch examples of neural data.

6.1.1 Neural data from the moving dots task

In addition to the famous orientation-sensitive neurons in visual cortex discovered by Hubel and Weisel [114], the medial temporal (MT) area contains motion sensitive neurons tuned to particular directions. Consider a pair of such cells, one preferentially responsive to right-going and the other to left-going motions, subject to the moving dots stimulus. The width of the tuning curves implies that the firing rates of the two cells may not differ much on average, and, especially for low coherence displays, they will be noisy, implying that decisions based on instantaneous activities of MT neurons would be inaccurate. Accuracy can be improved by integrating noisy signals and comparing the relative levels of the integrated quantities (this is basically the content of the optimal detection procedure of the sequential probability ratio test described below). The lateral interparietal area (LIP) and frontal eye fields (FEF), involved in eye movement control, have been proposed as sites for such integration.

Fig. 81 illustrates this idea via representations of typical firing rates observed in areas MT and LIP of monkeys trained on the moving dots task. Note that while it is hard to distinguish which MT signal dominates by direct observation, the (putatively) integrated LIP signals draw apart clearly as time elapses. In fact, after a transient due to the initial bursts of MT spikes, the activity of one population (black) *declines* while the other (gray) grows. This provides evidence for mutual inhibition, as in the model described in the next section.

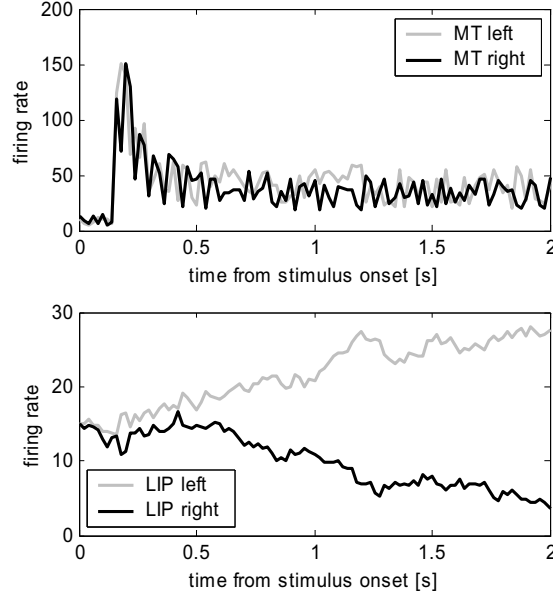


Figure 81: Cartoons of neural firing rates vs time from MT (top) and LIP (bottom) neurons sensitive to left-going (gray) and right-going (black) stimuli, for a left-going stimulus. This figure, taken from [17], is redrawn schematically based on data from [24, 186, 194].

6.1.2 A neural network model for 2AFC: leaky competing accumulators

Following [214, 17] and others, we model the accumulation of evidence in 2AFC tasks in LIP by two competing and mutually inhibitory neural populations, each selectively responsive to sensory input corresponding to one of the two alternatives. We already met a deterministic version of this *leaky accumulator model* (LAM) or *leaky competing accumulator model* (LCA) in Example 7 and Exercises 4-5 of §2.4.3. We now describe it in more detail. We consider a specific version of the LAM which may be written as

$$\tau \dot{x}_1 = -x_1 - \gamma f_{g,\beta}(x_2) + I_1 + c \eta_1(t), \quad (6.1.1a)$$

$$\tau \dot{x}_2 = -x_2 - \gamma f_{g,\beta}(x_1) + I_2 + c \eta_2(t) \quad (6.1.1b)$$

(cf. [25]), in which the state variables $x_j(t)$ denote mean input currents to cells of the j th neural population and the integration implicit in the differential equations models temporal summation of dendritic synaptic inputs [86]. The parameter γ sets the strength of mutual inhibition via population firing rates $f_{g,\beta}(x_j(t))$, where $f_{g,\beta}$ denotes the sigmoidal activation (“current-frequency” or “input-output”) function:

$$f_{g,\beta}(x) = \frac{1}{1 + \exp(-4g(x - \beta))} = \frac{1}{2} [1 + \tanh(2g(x - \beta))], \quad (6.1.2)$$

which has maximal slope g at $x = \beta$ (see Fig. 32 in §2.4.3). The stimulus received by each population has mean I_j and is polluted by noise $\eta_j(t)$ of strength c , the $\eta_j(t)$ being independent, identically distributed (i.i.d.) white noise processes with zero mean $\mathbb{E}[\eta_j(t)] = 0$ and variance unity. The time constant τ sets the rate at which neural activities decay in the absence of inputs. In the context of the moving dots task described above, the noisy inputs come (primarily) from left- and

right-sensitive MT cells, and the firing rates $f_{g,\beta}(x_j(t))$ describe activities of the corresponding LIP “integrator cells.” The noise can also represent other unmodeled inputs to LIP.

In free-response mode the decision is made and response initiated when the firing rate $f_{g,\beta}(x_j(t))$ of either population first exceeds a preset threshold θ_j , it being usually assumed that $\theta_1 = \theta_2$. For cued responses, the population with greatest firing rate at time t determines the decision. It is also generally assumed that activities x_j decay to zero after response and before the next trial, so that initial conditions for (6.1.1) are $x_j(0) = 0$. See [25] for further details, discussions of variable gain $g = g(t)$, and a second model in which population firing rates themselves are the state variables.

The LAM (6.1.1) is a nonlinear ODE with additive random inputs: an example of a *stochastic differential equation* (SDE). Explicit results are hard to come by for nonlinear stochastic systems, so we start our analysis by linearizing the sigmoidal response functions at the maximum gain point $x = \beta$. A (partial) justification for this can be found in the proposal of Cohen et al. [44] that neural circuits should “equilibrate” to work near this point to utilize the resulting sensitivity and avoid the lower and upper bounds on firing rates (Fig. 32). We therefore replace the functions $f_{g,\beta}(x_j(t))$ of (6.1.1) by $g x_j$ to obtain:

$$dx_1 = (-kx_1 - wx_2 + s_1) dt + c dW_1, \quad (6.1.3a)$$

$$dx_2 = (-kx_2 - wx_1 + s_2) dt + c dW_2. \quad (6.1.3b)$$

Here we have defined new parameters $s_j = I_j/\tau$, $k = 1/\tau$, $w = g\gamma/\tau$, the latter both positive, and written the LAM in Itô form [71] to emphasise the fact that white noise is a continuous, but not differentiable, function of time. Hence the terms $\eta_j = dW_j/dt$ in Eqn. (6.1.1) are not defined, and these SDEs are better written in the form (6.1.3) in which the noise terms are represented by Wiener processes which add independent random increments dW_j to each state variable during each time instant dt .

Definition 9. A Wiener process or Brownian motion on the interval $[0, T]$ is a random variable $W(t)$ that depends continuously on $t \in [0, T]$ and satisfies the following conditions [71, 99]:

- $W(0) = 0$.
- For $0 \leq s < t \leq T$, $W(t) - W(s) \sim \sqrt{t-s} N(0, 1)$, where $N(0, 1)$ is a normal distribution with zero mean and unit variance. (In words, the increments $W(t) - W(s)$ are normally distributed with zero mean and unit variance.)
- For $0 \leq s < t < u < v \leq T$, $W(t) - W(s)$ and $W(v) - W(u)$ are independent.

As noted in §4.2.2, in numerical simulations, each increment of the Wiener process is discretized for a timestep dt as follows [99]:

$$dW \sim \sqrt{dt} N(0, 1), \text{ so that } W(t + dt) = W(t) + \sqrt{dt} N(0, 1), \quad (6.1.4)$$

where samples are drawn independently from $N(0, 1)$ on each step. Because the normal distribution is used, the process is also referred to as *Gaussian*.

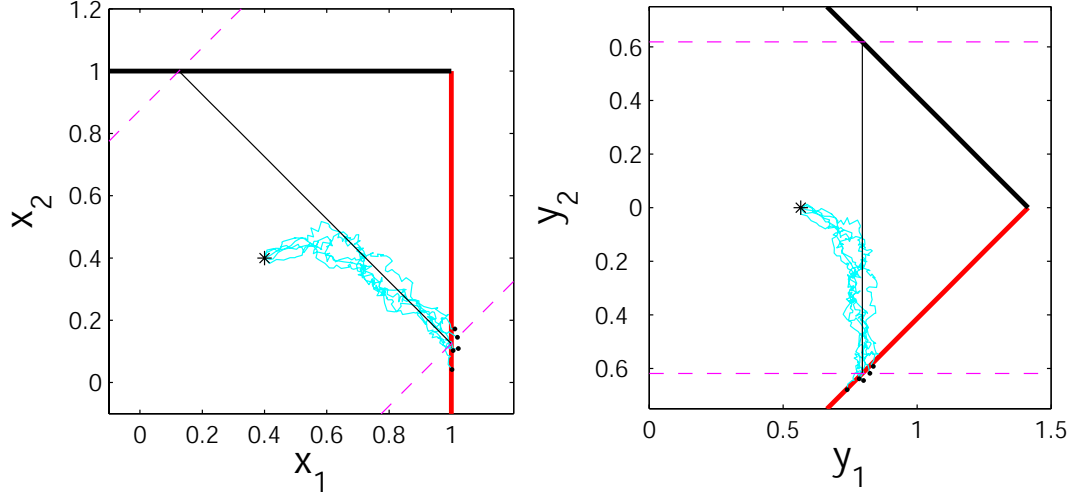


Figure 82: Five sample paths for the linearized LAM (6.1.3) showing the original and decoupled coordinate systems and illustrating convergence to an attracting line. The thick lines are the thresholds $x_i = \theta$; the attracting line is shown thin solid; the dashed lines are the thresholds $\pm\bar{\theta}$ for y_2 ; the star is the initial condition, and dots are final states following threshold crossing. Parameter values are $s_1 = 5$, $s_2 = 4$, $\theta = 1$, $c = 0.1$, $w = 4$, $k = 4$ (a balanced case). From [154, Fig. 1].

Making the orthonormal change of variables $y_1 = (x_1 + x_2)/\sqrt{2}$, $y_2 = (x_2 - x_1)/\sqrt{2}$ to an eigenvector basis, equations (6.1.3) decouple and become:

$$dy_1 = \left(-(w + k)y_1 + \frac{s_1 + s_2}{\sqrt{2}} \right) dt + c d\tilde{W}_1, \quad (6.1.5a)$$

$$dy_2 = \left(\lambda y_2 + \frac{s_2 - s_1}{\sqrt{2}} \right) dt + c d\tilde{W}_2, \quad (6.1.5b)$$

where $\lambda = w - k$ and the $d\tilde{W}_j$ are i.i.d. Wiener processes resulting from the orthogonal transformation. In the absence of noise ($c = 0$) the linear ODE (6.1.5a) is easily solved to reveal that $y_1(t)$ approaches a sink at $y_1^e = (s_1 + s_2)/[\sqrt{2}(w + k)]$. As we shall show in §6.4 below (Exercise 49), with noise present the sample paths of solutions of (6.1.5a) behave on average in a similar manner, approaching a normal distribution with mean $\mu = (s_1 + s_2)/[\sqrt{2}(w + k)]$ and variance $\sigma^2 = c^2/[2(w + k)]$ (a stochastic attractor). Thus, if $(w + k)$ is large, most sample paths of (6.1.5) converge to a neighborhood of an *attracting line* $y_1 = (s_1 + s_2)/\sqrt{2}(w + k)$ before crossing a threshold. Hence we may reduce to the scalar process (6.1.5b), with modified thresholds

$$\bar{\theta} = \pm \frac{2\theta(w + k) - (s_1 + s_2)}{\sqrt{2}(w + k)}. \quad (6.1.6)$$

Fig. 82 illustrates the behavior of typical solutions of the LAM and shows how the coordinate systems are related. After introducing further examples of SDEs, we shall provide a brief introduction to their analysis in §6.4 below.

Exercise 46. Verify the coordinate change based on eigenvectors from (6.1.3) to (6.1.5) and confirm that the transformed threshold value (6.1.6) is correct. Find analytical solutions of (6.1.3) for the noise-free ($c = 0$) case and compare then with numerical simulations of the original nonlinear system (6.1.1). Draw phase planes in both cases and explain how the differences influence threshold crossing behavior. [Remember to allow for the rescaling of time in passing from (6.1.1) to (6.1.3)].

In the linearized and reduced approximation of Eqn. (6.1.5b), it is only the *difference* in evidence that drives the decision process. Eqn. (6.1.5b) is an example of an Ornstein-Uhlenbeck (OU) process: a much studied linear stochastic differential equation. Furthermore, when $\lambda = w - k = 0$ we say that (6.1.3) is *balanced*, and (6.1.5b) becomes a drift-diffusion process (a Wiener process with constant drift rate). Quantitative comparisons showing fairly good agreements among the nonlinear, linearized, and reduced systems are given in [25]. However, note in Fig. 82 that, while solutions are attracted to the neighborhood of the line $y_1 = (s_1 + s_2)/[\sqrt{2}(w + k)]$, they can cross the original thresholds $x_1 = \theta$, $x_2 = \theta$ on either side of the points $y_2 = \pm\bar{\theta}$ for the reduced scalar process. The reduced OU or drift-diffusion process neglects nonlinear effects that limit sensitivity at high and low firing rates, *and* it approximates the threshold crossing dynamics for individual populations of left-and right-sensitive LIP cells by assuming “tight competition” (large values of leak and inhibition). See [17] for further details.

6.2 Derivation of firing rate equations from spiking neuron models

Before continuing to describe high level models of decision making, and in particular the optimal sequential probability ratio test and its continuum limit, the drift-diffusion process, we outline how a two-unit model similar to the leaky competing accumulators of §6.1.2 emerges by reduction of an integrate-and-fire model of interacting pools of excitatory and inhibitory cortical cells subject to stochastic spike inputs. Such models were first proposed to investigate working memory, and in particular to address the question of how activity that persists for several seconds can arise from individual neurons and synapses that have time constants of at most $\mathcal{O}(10^2)$ msec (NMDA synapses) [4, 31].

Building on these models, Wang [225] performed simulations on a circuit containing three populations of excitatory (pyramidal type) neurons, two of which are respectively sensitive to each of the two stimuli, the other being nonselective, and a single population of inhibitory neurons that globally suppress the excitatory cells. He showed that the firing rates of the two selective populations behaved in a manner (loosely) similar to the activity levels of the LAM network of Eqn. (6.1.1) [225] (specifically, compare Fig. 82 above with [225, Fig. 3B]). Subsequently, Wong and Wang [236] used mean field theory and some additional approximations to derive a reduced, nonlinear two-unit model, as we now describe.

The network in [236] contains 2000 leaky integrate-and-fire neurons divided into four groups: two stimulus-selective populations each containing 240 excitatory pyramidal cells, a non-selective pool of 1120 pyramidal cells, and an inhibitory population of 400 interneurons, connected as shown in Fig. 83(a). Recalling §4.2, The state variables are the cellular trans-membrane voltages $V_j(t)$, the internal synaptic variables $S_{AMPA,j}(t)$, $S_{NMDA,j}(t)$ and $S_{GABA,j}(t)$, and noisy external inputs $S_{AMPA,ext,j}(t)$ are applied to all 2000 cells.

The ODEs describing the subthreshold voltages and the fast synaptic dynamics are

$$C_j \frac{dV_j}{dt} = -g_L(V_j - V_L) + I_{syn,j}(t), \quad (6.2.1)$$

$$\frac{dS_{type,j}}{dt} = -\frac{S_{type,j}}{T_{type}} + \sum_l \delta(t - t_j^l), \quad (6.2.2)$$

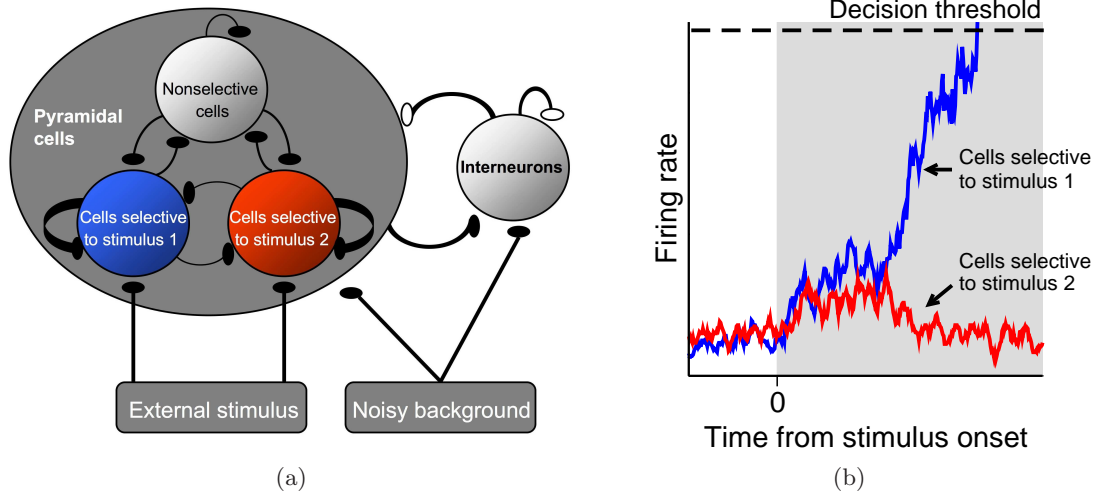


Figure 83: (a) The network of [236] contains three populations of excitatory cells; each selective population responds preferentially to one stimulus, the third is nonselective to both stimuli. A fourth population of interneurons provides overall inhibition. Excitatory (NMDA- and AMPA-mediated) and inhibitory (GABA_A-mediated) synapses are denoted by filled and open ovals respectively. All cells receive noisy AMPA-mediated background excitation; each cell connects to every other and selective populations have relatively stronger local recurrent excitation. (b) Stimuli excite both selective populations, but inhibition typically suppresses one population, producing winner-take-all dynamics. A decision is made when the first population crosses a fixed decision threshold. Figure adapted from [53].

where $type = AMPA, GABA, \text{ or } AMPA, ext$ and T_{type} is the time constant for that synapse type. Two ODEs are needed to describe the fast rise and slow fall of *NMDA*:

$$\frac{dS_{NMDA,j}}{dt} = -\frac{S_{NMDA,j}}{\tau_{NMDA,decay}} + \alpha x_j [1 - S_{NMDA,j}], \quad (6.2.3)$$

$$\frac{dx_j}{dt} = -\frac{x_j}{\tau_{NMDA,rise}} + \sum_l \delta(t - t_j^l). \quad (6.2.4)$$

When $V_j(t)$ crosses a threshold V_{thresh} at time t_j^l the cell emits a delta function $\delta(t - t_j^l)$, after which V_j is instantaneously reset and held at V_{reset} for an absolute refractory period τ_{ref} . The sums over delta functions in (6.2.2) and (6.2.4) represents spikes in presynaptic neurons, and, in the absence of stimuli, the external input currents are generated by Gaussian noise of mean μ and standard deviation σ :

$$dS_{AMPA,ext,j} = -(S_{AMPA,ext,j} - \mu_j) \frac{dt}{\tau_{AMPA}} + \sigma_j \sqrt{\frac{dt}{\tau_{AMPA}}} \mathcal{N}(0, 1). \quad (6.2.5)$$

In Eqns. (6.2.1-6.2.5) subscripts j index individual cells, and superscripts l index the times t_j^l at which the j th cell emits spikes. These external inputs are the sole source of randomness in the system.

The synaptic currents in Eqn. (6.2.1) are given in terms of the synaptic variables as follows:

$$\begin{aligned}
I_{syn,j}(t) &= I_{AMPA,ext,j}(t) + I_{AMPA,rec,j}(t) + I_{NMDA,rec,j}(t) + I_{GABA,rec,j}(t), \\
I_{AMPA,ext,j}(t) &= -g_{AMPA,ext,j}(V_j - V_E)S_{AMPA,ext,j}(t), \\
I_{AMPA,rec,j}(t) &= -g_{AMPA,rec,j}(V_j - V_E) \sum_{k=1}^{N_E} w_{k,j} S_{AMPA,k}(t), \\
I_{NMDA,rec,j}(t) &= -\frac{g_{NMDA,rec,j}(V_j - V_E)}{(1 + [Mg^{2+}] \exp(-0.062V_k)/3.57)} \sum_{k=1}^{N_E} w_{k,j} S_{NMDA,k}(t), \\
I_{GABA,rec,j}(t) &= -g_{GABA,rec,j}(V_j - V_I) \sum_{k=1}^{N_I} w_{I,j} S_{GABA,k}(t).
\end{aligned} \tag{6.2.6}$$

Here N_E and N_I denote the numbers of excitatory and inhibitory cells (1600 and 400, for the simulations of [236]), the subscript *rec* (henceforth omitted) indicates recurrent connections within the network, V_E and V_I are the glutamatergic (excitatory) and GABAergic (inhibitory) reversal potentials, $w_{k,j}$ denotes the strength of the synaptic connection from population k to population j , $g_{type,j}$ are synaptic conductances and g_L in Eqn. (6.2.1) denotes the leak conductance. For each synapse type there are two conductance values: $g_{type,p}$ for post-synaptic pyramidal neurons and $g_{type,I}$ for post-synaptic interneurons. Recurrent excitatory connections within the selective populations 1 and 2 have strength $w_{1,1} = w_{2,2} = w_+$, which are normally set greater than all other connections within and between the excitatory and inhibitory populations.

Stimuli are represented by the addition of terms $\mu_0(1 \pm c'/100)$ to the mean inputs μ_1 and μ_2 to the two populations of selective cells, with appropriate adjustments to the variances σ_j . Here μ_0 is the overall strength of the stimulus and c' denotes its discriminability: specifically, the percentage of coherently moving dots in the random motion display. (More generally, $c'/100 = 1$ for perfectly clear stimuli with infinite signal-to-noise ratio and $c'/100 = 0$ for zero SNR.) To run simulations of the 2AFC task, the ODEs are given random (small) initial conditions and integrated for a “fixation period” without stimuli, after which the stimuli $\mu_0(1 \pm c'/100)$ are applied and integration continues until the collective firing rate of any of the three excitatory populations exceeds a fixed decision threshold: see Fig. 83(b). If $c' > 0$ (resp $c' < 0$) and population 1 (resp. 2) first crosses threshold, then the response is correct; otherwise, it is an error, and if the threshold is crossed prior to stimulus onset, a premature response is logged. Further details, along with parameter values appropriate for modeling area LIP, appear in [236] and the supplementary materials to that paper; also see [53, 54].

Using the all-to-all coupling structure and eliminating irrelevant $S_{type,j}$ ’s (excitatory neurons have no *GABA*-synapses, inhibitory neurons have no *AMPA*- or *NMDA*-synapses), Eqns. (6.2.1-6.2.5) still constitute a 9200-dimensional stochastic dynamical system that is analytically intractable and computationally intensive to simulate. Following [236] we now sketch a sequence of low-dimensional reductions that preserve key physiological detail, permit bifurcation analyses, and relate the spiking network to the leaky accumulators described in §6.1.2.

We first reduce to a four-population model using a mean field approach from statistical physics [236], simplifying the self-consistency calculations of [31, 176] by employing a fixed average voltage $\bar{V} = (V_{reset} + V_{thresh})/2$ to estimate synaptic currents that enter each of the four cell populations via terms $J_{type,j} = -g_{type,j}(\bar{V} - V_{type})$. These are then multiplied by the appropriate number N_E or N_I

of presynaptic cells in each population and by an averaged synaptic variable $S_{type,j}$, and summed to create the incoming synaptic input currents to each population. Each term in the current to postsynaptic population j therefore takes the form

$$I_{type,j}(t) = N_{E/I} J_{type,j} \omega_{k,j} S_{type,k}(t). \quad (6.2.7)$$

Individual neuron voltages are replaced by averaged firing rates determined by frequency-current ($f - I$) relationships, analogous to the input-output function $f(\cdot)$ of Eqn. (6.1.2) (see §4.2.2 above). This yields an 11-dimensional system described by 4 firing rates $\nu_j(t)$, one inhibitory population-averaged synaptic variable $S_{GABA}(t)$, and two such variables $S_{AMPA,j}(t)$ and $S_{NMDA,j}(t)$ for each excitatory population (6 in all). The seven synaptic equations take the forms¹⁴

$$\frac{dS_{NMDA,j}}{dt} = -\frac{S_{NMDA,j}}{T_{NMDA}} + 0.641(1 - S_{NMDA,j})\nu_j, \quad (6.2.8)$$

$$\frac{dS_{AMPA,j}}{dt} = -\frac{S_{AMPA,j}}{T_{AMPA}} + \nu_j, \quad (6.2.9)$$

$$\frac{dS_{GABA,I}}{dt} = -\frac{S_{GABA,I}}{T_{GABA}} + \nu_I, \quad (6.2.10)$$

where $j = 1, 2, 3$ for the excitatory populations, $T_{NMDA} = T_{NMDA,decay}$, and the four firing rates obey

$$\frac{d\nu_j}{dt} = \frac{-(\nu_j - \phi_j(I_{syn,j}))}{T_j}, \quad (6.2.11)$$

with $j = 1, 2, 3$ and $j = I$ for the inhibitory population. Thus, the four population model is described by 11 differential equations: a remarkable reduction from the 9,200 ODEs and SDEs of the full network.¹⁵

The $f - I$ relationships $\phi_j(I_{syn})$ of Eqn. (6.2.11) for the populations 1, 2, 3 and I are estimated by computing the firing rate of a leaky integrate-and-fire neuron (cf. §4.2.2) from the first passage time of the Ornstein-Uhlenbeck process (6.2.1), with additive Gaussian noise, across a single threshold [177, 6, 176]:

$$\phi(I_{syn}) = \left[\tau_{ref} + \tau_m \sqrt{\pi} \int_{\frac{V_{reset}-V_{ss}}{\sigma_V}}^{\frac{V_{thresh}-V_{ss}}{\sigma_V}} e^{x^2} [1 + \operatorname{erf}(x)] dx \right]^{-1}, \quad (6.2.12)$$

where $\tau_m = C_j/g_L$, $V_{ss} = V_L + I_{syn}/g_L$ and σ_V is the standard deviation in the membrane potential due to the current fluctuations. The integral of (6.2.12) cannot be evaluated explicitly, and the following approximation, due to [1], was employed in [236]:

$$\phi(I_{syn}) \approx \frac{c_{E/I}(I_{syn} - I_{E/I})}{1 - \exp[-\gamma_{E/I}(c_{E/I}(I_{syn} - I_{E/I}))]}. \quad (6.2.13)$$

In Eqn. (6.2.13) the subscripts E/I denote excitatory and inhibitory neurons, I_{syn} is the total synaptic input to a single cell, $c_{E/I}$ is a gain factor and $\gamma_{E/I}$ determined how sharply the $f - I$

¹⁴The factor 0.641 in Eqn. (6.2.8) appears as a function value $F(\psi(r_i))$ in [236, Eqn. (6)] corresponding to a steady state firing rate, which appears in a subsequent approximation step below.

¹⁵The number 7200 that appears in [236, Fig. 1] evidently excludes the 2000 uncoupled SDEs (6.2.5) that determine the external inputs. Indeed, since these are stable OU processes one can in fact draw independent samples from the equilibrium distribution of sample paths for the OU process (see §6.4.2, below).

curve turns from $\phi(I_{syn}) \approx 0$ at low currents to an approximately linear relationship $\phi(I_{syn}) \approx c_{E/I}(I_{syn} - I_{E/I})$ at high currents. In [236] the parameters $c_{E/I}$, $\gamma_{E/I}$ and $I_{E/I}$ are chosen separately for the excitatory and inhibitory neurons, and the authors remark that the fit to Eqn. (6.2.12) is accurate for large input noise and moderate firing rates. This is appropriate for the decision making application, since decision thresholds are observed to lie at around 40 – 50 Hz.

The next simplification adopted in [236] uses the fact that, as observed in simulations of the full network, the firing rate of the nonselective cells $\nu_3(t)$ does not vary much across a trial. It is therefore assumed to remain constant, thereby removing the variables $S_{NMDA,3}$, $S_{AMPA,3}$ and ν_3 and replacing them by suitable constants. The firing rate of the inhibitory interneurons is also approximated by linearizing the $f - I$ curve ϕ_I for that population in the appropriate range of 8 – 15 Hz (again obtained from simulations):

$$\phi_I(I_{syn}) = \frac{c_I(I_{syn} - I_I)}{\gamma_2} + r_0, \quad (6.2.14)$$

with suitable choices for γ_2 and r_0 . This approximation makes it possible to compute a self-consistent firing rate for the inhibitory population in the presence of self-inhibition. The fact that the time constants for *GABA* and *AMPA* are short compared to that for *NMDA* ($T_{GABA} = 5$ ms, $T_{AMPA} = 2$ ms $\ll T_{NMDA} = 100$ ms) then justifies use of the quasi-equilibrium approximations $S_{GABA,I} = T_{GABA}\nu_I$ and $S_{AMPA,j} = T_{AMPA}\nu_j$, $j = 1, 2$ from Eqns. (6.2.9-6.2.10). (Recall that a similar approximation, based on separation of times scales [121, 89] was made in reducing the Hodgkin-Huxley system to two variables in §3.4.) Finally we assume that the time constants T_1, T_2 in Eqns. (6.2.11) are also much faster than T_{NMDA} , so that the firing rates also rapidly reach quasiequilibria: $\nu_j = \phi_j(I_{syn,j})$, $j = 1, 2$. Thus we set

$$S_{GABA,I} = T_{GABA}\nu_I = T_{GABA}\phi_I \text{ and } S_{AMPA,j} = T_{AMPA}\nu_j = T_{AMPA}\phi_j, \quad j = 1, 2, \quad (6.2.15)$$

and only $S_{NMDA,1}$ and $S_{NMDA,2}$ remain as dynamical variables, evolving according to Eqn. (6.2.8) with $\nu_j = \phi_j$:

$$\frac{dS_{NMDA,j}}{dt} = -\frac{S_{NMDA,j}}{T_{NMDA}} + 0.641(1 - S_{NMDA,j})\phi_j(I_{syn,j}), \quad j = 1, 2. \quad (6.2.16)$$

The synaptic currents $I_{syn,j}$ in Eqns. (6.2.16) are determined from Eqns. (6.2.6) via Eqn. (6.2.7), but to close this calculation the *GABA* input currents and *AMPA* and *NMDA* currents due to population 3 must be replaced by self-consistently computed quasiequilibrium values, since the quantities that determine them have been removed as variables. Replacing the subscripts *NMDA* and *AMPA* by *N* and *A* for short, we therefore have

$$\begin{aligned} I_{syn,1} &= J_{N,11}S_{N,1} - J_{N,21}S_{N,2} + J_{A,11}T_A\phi_1(I_{syn,1}) - J_{A,21}T_A\phi_2(I_{syn,2}) \\ &\quad + I_{sc,1} + I_{stim,1} + I_{noise,1}, \end{aligned} \quad (6.2.17)$$

$$\begin{aligned} I_{syn,2} &= -J_{N,12}S_{N,1} + J_{N,22}S_{N,2} - J_{A,12}T_A\phi_1(I_{syn,1}) + J_{A,22}T_A\phi_2(I_{syn,2}) \\ &\quad + I_{sc,2} + I_{stim,2} + I_{noise,2}, \end{aligned} \quad (6.2.18)$$

where $J_{N,ij}$ and $J_{A,ij}$ denote the net self ($i = j$) and cross ($i \neq j$) connection strengths between populations 1 and 2, $I_{sc,j}$, $j = 1, 2$ are the self-consistent internal currents due to *AMPA*, *GABA* and population 3, $I_{stim,j} = \mu(1 \pm E)$ are the stimuli, and $I_{noise,j}$ represent additive Gaussian

noise. The negative signs before the cross terms $J_{N,12}, J_{A,21}$ etc. in Eqns. (6.2.17-6.2.18) indicate that the net interaction between populations 1 and 2 is inhibitory, because global inhibition from interneurons excited by each population dominates that population's self excitation.

While equilibria of Eqns. (6.2.16) can be found by setting their right hand sides to zero, the fact that Eqns. (6.2.17-6.2.18) contain nonlinear terms of the form $\phi_1(I_{syn,j})$ implies that the $I_{syn,j}$'s would have to be solved for recursively at each step in the integration process to find a time-dependent solution. In [236] this was overcome by finding effective nonlinear $f - I$ relationships of the forms

$$\begin{aligned} \phi_1 &= H(y_1, y_2) \quad \text{and} \quad \phi_2 = H(y_2, y_1), \quad \text{where} \\ y_1 &= J_{N,11}S_{N,1} - J_{N,21}S_{N,2} + I_{sc,1} + I_{stim,1} + I_{noise,1}, \\ y_2 &= -J_{N,12}S_{N,1} + J_{N,22}S_{N,2} + I_{sc,2} + I_{stim,2} + I_{noise,2}, \end{aligned} \tag{6.2.19}$$

as described in the supplementary materials to [236]. Finally, $\phi_1(I_{syn,1})$ and $\phi_2(I_{syn,2})$ in (6.2.16) were respectively replaced by $H(y_1, y_2)$ and $H(y_2, y_1)$, thus producing a well-defined pair of ODEs. From these one can derive bifurcation diagrams, indicating subcritical pitchforks and tristability, and associated phase portraits, as shown in [236, Figs. 4-5 and 12-13], samples of which are reproduced here in Figs. 84-85. The reader should compare the phase portraits of Fig. 84 with those of Example 7, Figs. 33-34 in §2.4.3.

Referring to Fig. 85, for small recurrent connection strengths w_+ (Region I) a unique branch of low amplitude stable states exists, but as w_+ increases a pair of asymmetric stable states appears over a small range of stimulus strengths μ_0 , bounded by pitchfork bifurcations (Region II). Branches of saddle-node bifurcations emerge from degenerate pitchforks, which become subcritical in Region III, producing ranges in μ_0 over which 5 equilibria exist, 3 stable and 2 unstable, as well as the bistable region illustrated in the phase portraits of Figs 84(b,c). For sufficiently high w_+ , after the order of the two pitchfork bifurcations changes on the μ_0 axis, 4 stable equilibria and 5 unstable equilibria coexist in the central region of the bifurcation diagrams of region IV, providing phase portraits that are considerably more complex than those of the Usher-McClelland model of Example 7 (not shown here in bifurcation diagram IV, but see the analogous diagrams in [54, Figs. 5-7, 11 and 13]).

The phase portrait of Fig. 84(a) has 3 stable equilibria, one with both $S_{N,1}$ and $S_{N,2}$ low, and the others with $S_{N,1} \gg S_{N,2}$ and $S_{N,2} \gg S_{N,1}$. We refer to these as the low-low, high-low and low-high attractors; their domains of attraction are separated by stable manifolds of the two saddle points that lie between them. The latter attractors represent working memory states that can store decisions for alternatives 1 and 2 during a delay period, while the low-low attractor corresponds to an undecided or pre-decision state. At stimulus onset the low-low state typically becomes unstable or disappears (see Figs. 84(b,c,d)), causing the solution to jump to the high-low or low-high attractor; when a stable low-low state persists, its domain of attraction shrinks and noise causes the jump. If thresholds are applied, they would cross the phase plane between the low-low and high-low and low-high states, e.g. at $S_{N,1} = 0.55$ and $S_{N,2} = 0.55$. In this context, the fourth attractor that occurs for large w_+ corresponds to a high-high state that lies above *both* thresholds and represents an impulsive choice. See [236] and [54] for further information and interpretations.

Note that the regions labeled “bistable” in Fig. 85(top), bounded by the saddle-node and pitchfork bifurcation curves, are in fact tristable since three stable states (low-low, low-high and high-low)

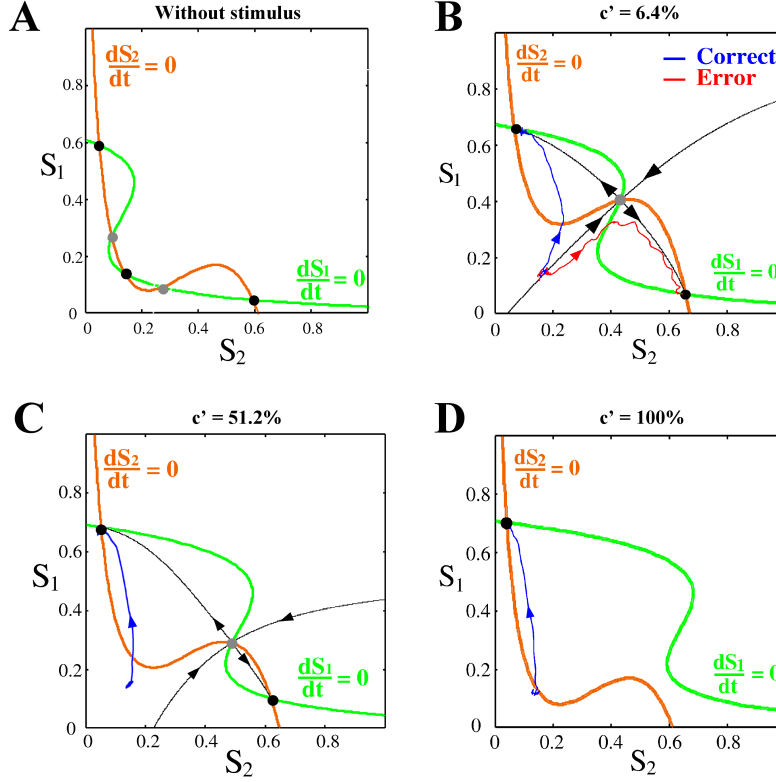


Figure 84: Phase portraits of the reduced noise-free system (6.2.16) with effective $f-I$ relationships (6.2.19); nullclines shown as bold red and green curves. (a): no stimulus, $\mu_0 = 0$; (b,c,d): $\mu_0 > 0$ and $c' = 6.4\%$ (b), 51.2% (c) and 100% (d). Note reflection-symmetry about $S_1 = S_2$ and tristability in (a); bistability and increasing symmetry breaking as c' increases through (b,c), and a unique global attractor in (d). In (b-d) the upper left sink corresponds to the correct choice, and blue and red sample paths in (b) respectively indicate correct and error choices for the system with small additive white noise. Reproduced from [236, Fig. 5], which uses the notation $S_j = S_{N,j}$ ($= S_{NMDA,j}$).

coexist, as indicated by bifurcation diagram III; the region labeled “competition” is bistable. See the phase portraits of Fig. 84(a) and (b) for examples.

More recently the effects of neuromodulation by norepinephrine have been studied using a modified version of the Wong-Wang model [53], and mean field reduction of this circuit to a four- and ultimately a two-population model has been carried out [54], although the final reduction to two variables as above was not employed, since effective relationships of the form (6.2.19) could not be found over the entire parameter space necessary to model neuromodulation. (Nonetheless, nullclines for 2-dimensional projected phase planes were computed and used to elucidate the dynamics, and bifurcation diagrams analogous to those of Fig. 85 were computed.) Local reductions of a more rigorous nature to 1-dimensional systems have also been done from similar models in the neighborhood of a pitchfork bifurcation [184], using center manifold theory. For a review of decision making and models of neural circuits, see [226].

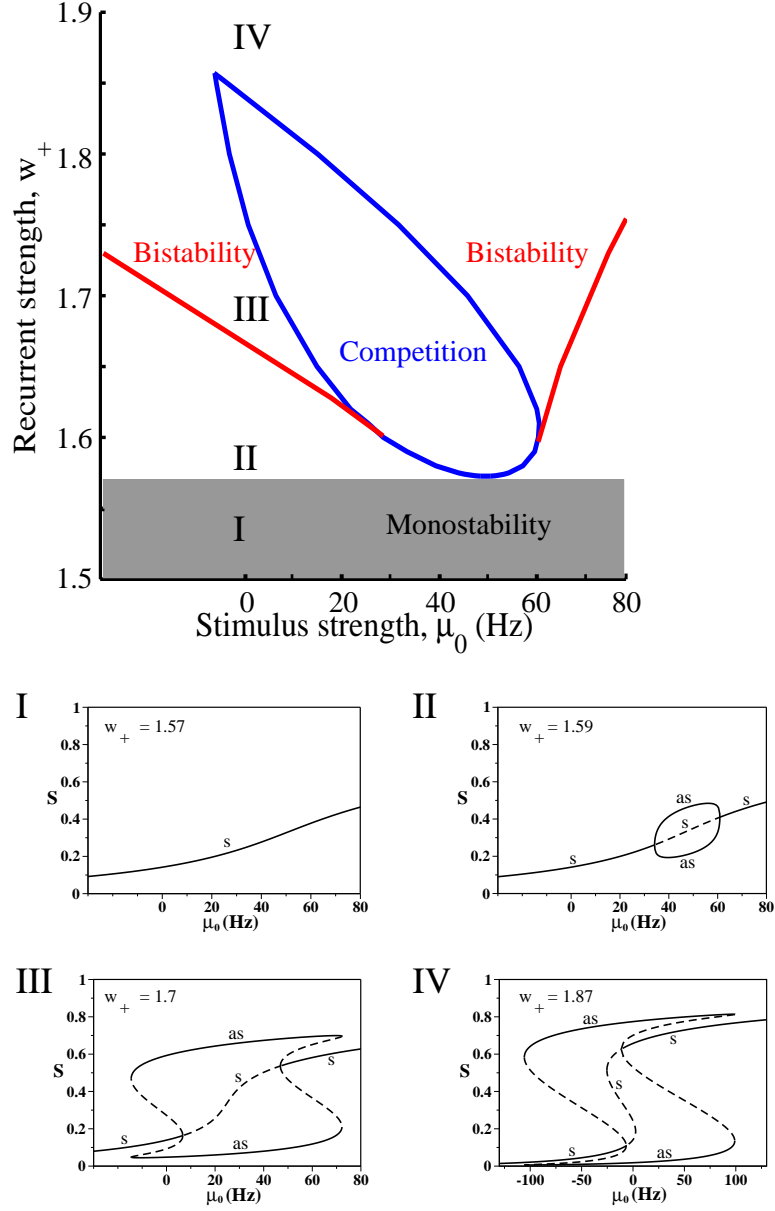


Figure 85: A bifurcation set (top) and bifurcation diagrams (I-IV below) for the reduced system (6.2.16) with effective $f - I$ relationships (6.2.19), as stimulus strength μ_0 and recurrent connection strength w_+ vary, with $c' = 0$; red and blue curves respectively denote saddle-node and pitchfork bifurcations. In (I-IV) S and AS respectively denote symmetric ($S_{N,1} = S_{N,2}$) and asymmetric ($S_{N,1} \neq S_{N,2}$) states, which occur in reflection-symmetric pairs for $c' = 0$. Solid and dashed curves respectively denote stable and unstable equilibria. Reproduced from [236, Fig. 12]. See text for discussion.

6.3 The best way to choose between two noisy signals

We now depart from the neural evidence and accumulator models for a while, to describe a discrete procedure from mathematical statistics and signal analysis that delivers decisions in binary

choice situations that are optimal in a precise sense. Remarkably, a continuum limit of this method turns out to be a drift-diffusion process of the type that emerged in §6.1.2 above. First we outline a standard procedure for testing samples of fixed size.

Suppose we wish to decide whether a random sequence $Y = y_1, y_2, \dots, y_N$ of N independent observations is drawn from the probability distribution $p_0(y)$ (hypothesis H_0) or $p_1(y)$ (hypothesis H_1). Neyman and Pearson [160] showed that the optimal procedure is to calculate the following ratio:

$$\frac{p_{1N}}{p_{0N}} = \frac{p_1(y_1)p_1(y_2) \cdots p_1(y_N)}{p_0(y_1)p_0(y_2) \cdots p_0(y_N)}, \quad (6.3.1)$$

and to accept hypothesis H_0 (resp., H_1) if $\frac{p_{1N}}{p_{0N}} < K$ (resp., $\frac{p_{1N}}{p_{0N}} \geq K$), where K is a constant determined by the desired level of accuracy for one of the hypotheses. Setting $K < 1$ increases the reliability of the decision to accept H_0 , at the expense of reducing the reliability of accepting H_1 . Since the p_{iN} , $i = 0, 1$ are the probabilities of Y occurring under the hypotheses H_i , setting $K = 1$ reduces the procedure to simply determining which hypothesis is most likely (determining *maximum likelihood*). In this case, if the hypotheses occur with equal probability, the procedure is optimal in that it guarantees the smallest overall error rate.

6.3.1 The sequential probability ratio test

The Neyman-Pearson procedure is optimal for fixed sequence sizes N . When this requirement is relaxed, the optimal procedure is the *sequential probability ratio test* (SPRT), in which observations continue as long as the running quotient $\frac{p_{1n}}{p_{0n}}$ (defined as in (6.3.1) with $N = n$) satisfies the inequality

$$A_0 < \frac{p_{1n}}{p_{0n}} < A_1, \quad (6.3.2)$$

where $A_0 < 1 < A_1$ are two given constant thresholds. The hypothesis H_0 (resp., H_1) is accepted at step n as soon as $\frac{p_{1n}}{p_{0n}} \leq A_0$ (resp., $\frac{p_{1n}}{p_{0n}} \geq A_1$). The SPRT was independently developed during World War II by Abraham Wald [218], who was introduced to the problem by Milton Friedman and W. Allen Wallis while all were members of the Statistical Research Group at Columbia University [221], and by George Barnard [12, 51] in the U.K. Alan Turing and his coworkers at Bletchley Park employed the SPRT to break the Enigma code used by the German navy in World War II [82, 80]¹⁶.

The SPRT is optimal in the following sense. Let $P(\text{rej } H_i | H_i)$ be the probability that hypothesis H_i is true but rejected, $i = 0, 1$, and let $E_i(N)$ be the expected value for the number of observations required to reach a decision when hypothesis H_i is true, $i = 0, 1$.

Theorem 5. *Among all fixed sample or sequential tests for which*

$$P(\text{rej } H_i | H_i) \leq \alpha_i, \quad i = 0, 1,$$

and for which $E_0(N)$ and $E_1(N)$ are finite, the SPRT with error probabilities $P(\text{rej } H_i | H_i) = \alpha_i$, $i = 0, 1$, minimizes both $E_0(N)$ and $E_1(N)$.

¹⁶Both Wald and Turing died prematurely: Wald in a plane crash en route to a scientific presentation in India in 1950, and Turing (apparently) committed suicide in 1954.

This theorem was first proved in [219]; for a simpler proof, see [140].

The thresholds A_0 and A_1 in the SPRT are related to the *error rates* α_0 and α_1 as follows [218, 140]. Consider the set C_1 of n -length sequences Y such that the SPRT chooses H_1 when Y occurs. That is, for any $Y \in C_1$,

$$p_1(y_1)p_1(y_2) \cdots p_1(y_n) \geq A_1 p_0(y_1)p_0(y_2) \cdots p_0(y_n).$$

Integrating this inequality over C_1 , we get

$$p_1(C_1) \geq A_1 p_0(C_1), \quad (6.3.3)$$

where $p_j(C_1)$ is the probability of making choice 1 given that hypothesis H_j is true. By definition, $p_1(C_1) \geq 1 - \alpha_1$ and $p_1(C_0) \leq \alpha_0$, so that

$$1 - \alpha_1 \geq A_1 \alpha_0 \quad \Rightarrow \quad A_1 \leq \frac{1 - \alpha_1}{\alpha_0}.$$

Similarly,

$$\alpha_1 \leq A_0(1 - \alpha_0) \quad \Rightarrow \quad A_0 \geq \frac{\alpha_1}{1 - \alpha_0}.$$

The inequalities fail to be equalities because it is possible to overshoot the boundaries A_0 or A_1 . However, in practice, there is typically little penalty in assuming equality [218, 140]:

$$A_0 = \frac{\alpha_1}{1 - \alpha_0}, \quad A_1 = \frac{1 - \alpha_1}{\alpha_0}. \quad (6.3.4)$$

Note that, when using an SPRT with A_0 and A_1 defined in this way, the condition that $A_0 < 1 < A_1$ in the proof becomes

$$\frac{\alpha_1}{1 - \alpha_0} < 1 < \frac{1 - \alpha_1}{\alpha_0}. \quad (6.3.5)$$

Thus, the proof requires that $\alpha_1 < 1 - \alpha_0$. Now, $1 - \alpha_0$ is the probability of choosing H_0 when H_0 is true. Thus, the proof requires that the probability of choosing H_0 when H_1 is true is less than the probability of choosing H_0 when H_0 is true. Similarly, it requires that $\alpha_0 < 1 - \alpha_1$, that is, the probability of choosing H_1 when H_0 is true is less than the probability of choosing H_1 when H_1 is true: a reasonable restriction for a decision making procedure. Wald [218] also gives approximate expressions for the expected numbers of observations that may be written

$$E_1(N) \approx \frac{\alpha_1 \log \left(\frac{\alpha_1}{1 - \alpha_0} \right) + (1 - \alpha_1) \log \left(\frac{1 - \alpha_1}{\alpha_0} \right)}{E_1 \left(\log \left(\frac{p_1(y)}{p_0(y)} \right) \right)} \quad (6.3.6)$$

$$E_0(N) \approx \frac{(1 - \alpha_0) \log \left(\frac{\alpha_1}{1 - \alpha_0} \right) + \alpha_0 \log \left(\frac{1 - \alpha_1}{\alpha_0} \right)}{E_0 \left(\log \left(\frac{p_1(y)}{p_0(y)} \right) \right)}, \quad (6.3.7)$$

where $E_i \left(\log \left(\frac{p_1(y)}{p_0(y)} \right) \right)$ is the expected value of the argument when H_i is true, $i = 0, 1$. In particular, when we set $\alpha_0 = \alpha_1 = \text{ER}$, so that accuracy is the same for both alternatives, Eqns. (6.3.6-6.3.7) become:

$$E(N) \approx \frac{(1 - 2\text{ER}) \log \left(\frac{1 - \text{ER}}{\text{ER}} \right)}{E \left(\log \left(\frac{p_1(y)}{p_0(y)} \right) \right)}. \quad (6.3.8)$$

6.3.2 Random walks and the continuum limit of SPRT

Laming models the 2AFC by supposing that decisions are made based on accumulation of information [136]. In each trial the subject makes a series of brief observations of the stimulus (\mathcal{S}_0 or \mathcal{S}_1) represented by the random sequence y_1, y_2, \dots, y_n . The increment of information gained from (independent) observation y_r is defined to be

$$\delta I_r = \log \left(\frac{p_1(y_r)}{p_0(y_r)} \right), \quad (6.3.9)$$

where $p_i(y)$ is the probability distribution for y given that stimulus \mathcal{S}_i was presented, $i = 0, 1$. (Implicitly, the subject has some internal representation of $p_0(y)$ and $p_1(y)$.) At the n^{th} observation (cf. Eqn. (6.3.1)), the total information accumulated is

$$I_n = \sum_{r=1}^n \delta I_r = \sum_{r=1}^n \log \left(\frac{p_1(y_r)}{p_0(y_r)} \right) = \log \left(\frac{p_{1n}}{p_{0n}} \right) : \quad (6.3.10)$$

this is the *log likelihood ratio* [78]. Under the free response protocol observations continue as long as $\mathcal{I}_0 < I_n < \mathcal{I}_1$, where \mathcal{I}_0 and \mathcal{I}_1 are the thresholds in the logarithmic space, and a choice is made at step n if $I_n \leq \mathcal{I}_0$ or $I_n \geq \mathcal{I}_1$. Hence this log likelihood formulation is equivalent to making decisions using the SPRT with $\mathcal{I}_0 = \log A_0$ and $\mathcal{I}_1 = \log A_1$. For example, if the desired error rates are $\alpha_0 = \alpha_1 = \epsilon$, which is reasonable if the signals \mathcal{S}_0 and \mathcal{S}_1 are equally salient, from (6.3.4) we take

$$\mathcal{I}_0 = \log \left(\frac{\epsilon}{1-\epsilon} \right) < 0, \quad \mathcal{I}_1 = \log \left(\frac{1-\epsilon}{\epsilon} \right) = -\mathcal{I}_0 > 0, \quad (6.3.11)$$

cf. [136]. (The signs follow from the assumed inequality (6.3.5).)

Thus, from (6.3.10), in logarithmic variables the trajectory I_n is a discrete-time, biased random walk with initial condition zero: a new increment of information arrives, and the trajectory is updated, as the timestep advances from $n \rightarrow n+1$ (recall that the increments δI_r are assumed to be independent and identically distributed). Hereafter we treat the continuous-time limit $I(t)$ of this process, in which infinitesimal increments of information arrive at each moment in time. This limit must be taken with some care in order to preserve the variability present in (6.3.10). Up to an unimportant scale factor between timesteps n and the continuous time t , the limiting procedure is as follows. Let the δI_r have mean m and variance D^2 (assumed finite). Then define the family (indexed by $N = 1, 2, \dots$) of random functions of $t \in [0, T]$, where T is some large time, as follows:

$$I^N(t) = \frac{1}{\sqrt{N}} \sum_{r=1}^k (\delta I_r - m) + \frac{1}{N} \sum_{r=1}^k \delta I_r \quad \text{where } k = \lfloor Nt/T \rfloor. \quad (6.3.12)$$

Here, $\lfloor Nt/T \rfloor$ is the largest integer smaller than Nt/T . Note that the first term of (6.3.12) is normalized by $1/\sqrt{N}$ and the second by $1/N$, reflecting the different rates at which fluctuations and means accumulate as random increments are summed. For *any* N , $I^N(t)$ has mean $m \lfloor t/T \rfloor$ and variance $D^2 \lfloor t/T \rfloor$; e.g., from (6.3.10), I_n has mean mn and variance $D^2 n$. Furthermore, the Donsker Invariance Principle (see [16, Thm. 37.8]), together with the Law of Large Numbers, implies that as $N \rightarrow \infty$

$$I^N(t) \Rightarrow DW(t) + mt \equiv I(t), \quad (6.3.13)$$

where $W(\cdot)$ is a Wiener process and the random functions $I^N(\cdot)$ converge in the sense of distributions.

Eqn. (6.3.13) implies that the limiting process $I(t)$ satisfies the stochastic differential equation (SDE)

$$dI = m dt + D dW, \quad I(0) = 0, \quad (6.3.14)$$

with thresholds $\mathcal{I}_0 < 0 < \mathcal{I}_1$. The drift m and variance D of the δI_r and hence of (6.3.14) depend upon the distributions $p_i(y)$, cf. (6.3.9). For example, in the case of Gaussians

$$p_0(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu_0)^2/(2\sigma^2)}, \quad p_1(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu_1)^2/(2\sigma^2)}, \quad (6.3.15)$$

with $\mu_1 > \mu_0$, we have

$$\delta I_r = \log \left(\frac{p_1(y_r)}{p_0(y_r)} \right) = \frac{\mu_1 - \mu_0}{\sigma^2} \left(y_r - \frac{\mu_0 + \mu_1}{2} \right), \quad (6.3.16)$$

and if \mathcal{S}_i is presented, the expected value of y_r is $E(y_r) = \mu_i$, and the variance is $Var(y_r) = \sigma^2$. Thus, taking expectations and substituting in (6.3.16), we obtain

$$E(\delta I_r) = \pm \frac{(\mu_1 - \mu_0)^2}{2\sigma^2} = m, \quad (6.3.17)$$

(the $+$ applies if \mathcal{S}_1 is presented, the $-$ if \mathcal{S}_0), and in both cases

$$Var(\delta I_r) = D^2 = \frac{(\mu_1 - \mu_0)^2}{\sigma^2}, \quad (6.3.18)$$

cf. [79, 80]. Thus the limiting SDE is

$$dI = \frac{\mu_1 - \mu_0}{\sigma^2} \left[\left(\frac{\mu_1 - \mu_0}{2} \right) dt + \sigma dW \right], \quad I(0) = 0. \quad (6.3.19)$$

If each incremental observation δI_r is composed of many subobservations (for example, from different regions of the visual field, or from large populations of neurons), this Gaussian assumption is justified by the Central Limit Theorem.

In the particular case of (6.3.16) in which $\mu_1 = -\mu_0 = A$, $\sigma = c$, appropriate to tasks such as the moving dots task in which the alternative stimuli are of equal coherence, the simplified form of (6.3.16) implies that the accumulating information I_n is simply a scaled version of the running total of observations y_r :

$$\delta I_r = \frac{2A}{c^2} y_r \Rightarrow I_n = \sum_{r=1}^n \delta I_r = \frac{2A}{c^2} \sum_{r=1}^n y_r \stackrel{\text{def}}{=} \frac{2A}{c^2} y_n. \quad (6.3.20)$$

Assuming without loss of generality that \mathcal{S}_1 is presented, the y_r have mean A and variance c^2 , so that, in the continuous time limit analogous to (6.3.12), y_n converges to $y(t)$, which satisfies the SDE:

$$dy = \pm A dt + c dW; \quad y(0) = 0. \quad (6.3.21)$$

The “logarithmic” SPRT involving observations δI_r is therefore equivalent to solving the first passage problem defined by (6.3.21) with thresholds $y = \pm z = \pm \frac{c^2}{2A} \mathcal{I}_1$, as in (6.3.11). We shall

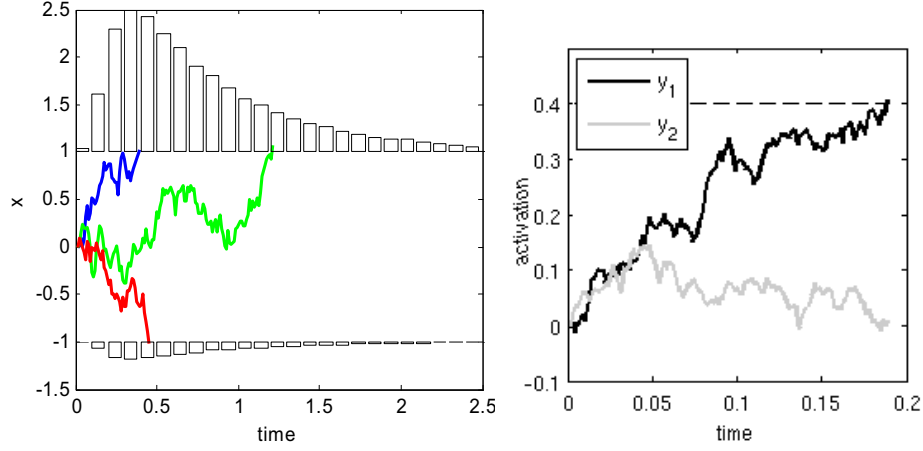


Figure 86: Left: three sample paths of the drift-diffusion process (6.3.21) along with histograms of first passage times for correct and incorrect responses obtained from 100,000 trials. Parameter values $A = c = z = 1$. Right: typical sample paths for the states x_j of the leaky accumulator model of §6.1.2. From [17, Fig. 2]. Sample paths were computed by the Euler-Maruyama method with step size $\Delta t = 0.01$ (§6.4.4). **NOTE: Interpret $x \mapsto y$ on left panel and $y_j \mapsto x_j$ on right.**

refer to (6.3.21) as a drift-diffusion (DD) process or drift-diffusion model (DDM). In computing first passage times and error rates below, we may appeal to reflection symmetry about $y = 0$ and assume without loss of generality that $A \geq 0$ so that stimulus \mathcal{S}_i is present and alternative 1 is correct.

Fig. 86 illustrates typical sample paths of the DDM and the resulting first passage time histograms that emerge from threshold crossing for both correct and incorrect responses. It also shows the states of the LAM of §6.1.2 to illustrate that the difference $x_2(t) - x_1(t)$ behaves like $y(t)$, after an initial transient during which both values rise at similar rates (cf. the phase portraits of Fig. 82). The constant drift SDE (6.3.14) or (6.3.21) is a particular limit of the discrete random walk occurring in the SPRT and Neyman-Pearson tests. In the following sections we analyse these stochastic processes in both free response and time-constrained contexts. First we need some facts about SDEs that were promised in §6.1.2.

6.4 Introduction to stochastic differential equations

In this section we discuss scalar SDEs; specifically, initial value problems of the form

$$dx = f(x, t) dt + g(x, t) dW, \quad x(t_0) = x_0, \quad (6.4.1)$$

which include the constant drift-diffusion process (6.3.21) derived above from the SPRT, and the Ornstein-Uhlenbeck (OU) processes (6.1.5) that emerged from the linearized LAM model of §6.1.2. Multidimensional generalizations also exist, but in practice explicit results can typically be obtained only in the scalar case, and often only for linear SDEs such as Eqns. (6.1.5).

In Eqn. (6.4.1) $f(x, t)$ and $g(x, t)$ are respectively called the drift and diffusion terms (or coefficients), and dW denotes Wiener increments (Definition 9, §6.1.2). We may rewrite the SDE in

integral form:

$$x(t) = x_0 + \int_{t_0}^t f(x(s), s) dt + \int_{t_0}^t g(x, s) dW(s), \quad (6.4.2)$$

where the first integral is the usual one from calculus which (implicitly) defines the unique solution of the deterministic ODE $\dot{x} = f(x, t)$. The second, stochastic integral can be interpreted as the continuum limit of a discrete sum, much as in the Riemann sum used in Euler's method (§§2.3.1-2.3.2; Fig. 22):

$$I_{\text{stoch}} = \sum_{j=0}^N g(x(t_j), t_j) dW(t_j) = \sum_{j=1}^N g(x(\tau_j), \tau_j) [W(t_j) - W(t_{j-1})], \text{ with } \tau_j \in [t_{j-1}, t_j]. \quad (6.4.3)$$

The sum (6.4.3) approaches the second integral of (6.4.2) as $\delta t \rightarrow 0$ and $N \rightarrow \infty$. (For simplicity we may take equal increments $t_j - t_{j-1} = \delta t$ and choose $\delta t = (t - t_0)/N$.) However, unlike the deterministic case, the choice of time instants τ_j at which the function $g(x(t), t)$ is evaluated can affect the result. We shall adopt the choice $\tau_j = t_{j-1}$, which corresponds to the Itô stochastic integral; (choosing $\tau_j = (t_j - t_{j-1})/2$ yields the Stratonovich integral) [71, §4.2], cf. [61, §10.1]. If the function g in Eqn. (6.4.3) is constant (i.e., independent of x and t , as for many examples considered below), then the choice of τ_j makes no difference.

Solutions of SDEs are random processes, and so we often want to average over many realizations to obtain probabilistic descriptions such as the probability density of solutions of (6.4.1). To derive these, we need an important result from Itô's stochastic calculus.

6.4.1 Change of variables: Itô's lemma

A primary difference between regular calculus and stochastic calculus occurs in the chain rules for differentiation. Recall from calculus that if $y(t) = h(x(t))$ for scalar functions, then $\frac{dy}{dt} = \frac{dh}{dx} \frac{dx}{dt}$. Itô's lemma, which we will formally derive below, shows that additional higher order terms are present in the analogue of the chain rule for stochastic calculus.

We shall illustrate the stochastic chain rule in connection with the SDE (6.4.1). Recall the following property of a Wiener process

$$dW = W(t + dt) - W(t) = \sqrt{dt} N(0, 1), \quad (6.4.4)$$

implied by the second condition of Definition 9 (§6.1.2), and which in turn implies that

$$dW = \mathcal{O}(\sqrt{dt}), \quad dW^2 = \mathcal{O}(dt) \text{ and } \mathbb{E}[(dW)^2] = dt. \quad (6.4.5)$$

Given a twice-differentiable function $h(x(t))$, we ask what SDE does h obey? Expanding to second order in a Taylor series we find:

$$\begin{aligned} dh(x) &= h(x + dx) - h(x) = h'(x)dx + \frac{1}{2}h''(x)dx^2 + \dots \\ &= h'(x)[f(x, t)dt + g(x, t)dW] + \frac{1}{2}h''(x)[f(x, t)dt + g(x, t)dW]^2 + \dots \\ &= h'(x)f(x, t)dt + h'(x)g(x, t)dW + \mathcal{O}(dt^{3/2}) + \frac{1}{2}h''(x)g^2(x, t)dW^2. \end{aligned}$$

In the third step we substitute (6.4.1) and then use (6.4.5) to estimate orders of magnitude of dW and $(dW)^2$ in the fourth step. Finally, we let $dt \rightarrow 0$ and appeal to the fact that $dW^2 \rightarrow \mathbb{E}[(dW)^2] = dt$ as $dt \rightarrow 0$ in Itô calculus (this is essentially due to the appearance of dW only in stochastic integrals like (6.4.2): see [71, §§4.2.5 and 4.3.3]). This yields **Itô's lemma**:

$$dh(x(t)) = \left[h'(x(t))f(x(t), t) + \frac{1}{2}h''(x(t))g^2(x(t), t) \right] dt + h'(x(t))g(x(t), t) dW(t), \quad (6.4.6)$$

also known as Itô's formula or Itô's rule. Note that, unlike the standard chain rule, an additional term $\frac{1}{2}h''(x)g^2(x, t)$ appears in the $\mathcal{O}(dt)$ deterministic part of dh (unless h is linear in x).

6.4.2 The forward Kolmogorov or Fokker-Planck equation

Since individual solutions or *sample paths* of an SDE are determined by successive increments $dW(t)$ drawn randomly from a distribution, we often consider *ensembles* of solutions produced by integrating the SDE many times with different i.i.d. (independent, identically distributed) sequences $dW(t)$. A reasonable question is then: Given an initial condition $x(t_0) = x_0$ at time t_0 , what is the probability $p(x, t|x_0, t_0)$ of finding the solution at a point x at time $t > t_0$? This can, of course, be estimated by repeated (Monte-Carlo) simulations of the SDE, but there is a more elegant answer.

It turns out that the conditional probability density $p(x, t|x_0, t_0)$ evolves according to a linear PDE. This can be derived from the SDE (6.4.1)

$$dx = f(x, t) dt + g(x, t) dW, \quad x(t_0) = x_0, \quad (6.4.7)$$

using Itô's lemma, the statement (6.4.6) of which we also repeat for convenience:

$$dh(x(t)) = \left[h'(x(t))f(x(t), t) + \frac{1}{2}h''(x(t))g^2(x(t), t) \right] dt + h'(x(t))g(x(t), t) dW(t). \quad (6.4.8)$$

Taking expectations of Eqn. (6.4.8) with respect to x , dividing by t , and using the fact that $\mathbb{E}[dW] = 0$, we obtain

$$\frac{d}{dt}\mathbb{E}[h(x)] = \mathbb{E}[h'(x)f(x, t)] + \frac{1}{2}\mathbb{E}[h''(x)g^2(x, t)]. \quad (6.4.9)$$

The expectations in Eqn. (6.4.9) are averages over x with respect to the joint probability density $p(x, t|x_0, t_0)$, i.e.

$$\mathbb{E}[h(x)] = \int_{-\infty}^{\infty} h(x)p(x, t|x_0, t_0) dx, \quad (6.4.10)$$

for any (continuous) function $h(x)$. Here we have assumed that $p(x, t|x_0, t_0)$ is defined on \mathbb{R} , as is appropriate for one-dimensional Brownian motion. Using (6.4.10) and writing $p(x, t|x_0, t_0) = p(x, t)$ for short, Eqn. (6.4.9) becomes

$$\begin{aligned} \frac{d}{dt} \int_{-\infty}^{\infty} h(x)p(x, t) dx &= \int_{-\infty}^{\infty} \left[h'(x)f(x, t) + \frac{1}{2}h''(x)g^2(x, t) \right] p(x, t) dx \\ &= \int_{-\infty}^{\infty} \left\{ [-f(x, t)p(x, t)]' + \frac{1}{2}[g^2(x, t)p(x, t)]'' \right\} h(x) dx. \end{aligned} \quad (6.4.11)$$

Here the second equality comes from integration by parts and using the fact that $p(x, t) \rightarrow 0$ and $p'(x, t) \rightarrow 0$ as $x \rightarrow \pm\infty$, conditions that are necessary for the probability density to be well-defined. Since Eqn. (6.4.11) must hold for arbitrary functions $h(x)$, the integrands of its left- and right-hand sides must be equal:

$$\frac{\partial}{\partial t}p(x, t) = -\frac{\partial}{\partial x}[f(x, t)p(x, t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[g^2(x, t)p(x, t)]. \quad (6.4.12)$$

This is called the *forward Fokker-Planck* or *Kolmogorov* equation for the SDE (6.4.7), reflecting its origins in both statistical physics and probability theory. See [71] for (much) more information.

To obtain a well-posed problem Eqn. (6.4.12) must be supplied with an initial condition in the form of a function

$$p(x, t_0) = p_0(x) \quad (6.4.13)$$

which encodes the distribution of starting points x_0 for sample paths of the SDE, and satisfies the normalization condition $\int_{-\infty}^{\infty} p(x, t_0) dx = 1$ for a probability distribution.

Exercise 47. *Verify that Eqn. (6.4.12) does preserve the value of the integral (mass) $\int_{-\infty}^{\infty} p(x, t) dx$ of the probability distribution. [This is an example of a Liapunov function for a PDE. The L^1 norm of $\int_{-\infty}^{\infty} p(x, t) dx$ remains constant.]*

Taking the simplest imaginable SDE

$$dx = \sigma dW, \quad x(0) = 0, \quad (6.4.14)$$

which is the continuum limit of a random walk, the corresponding Fokker-Planck equation is the classical diffusion or heat equation:

$$\frac{\partial p}{\partial t} = \frac{\sigma^2}{2} \frac{\partial^2 p}{\partial x^2}. \quad (6.4.15)$$

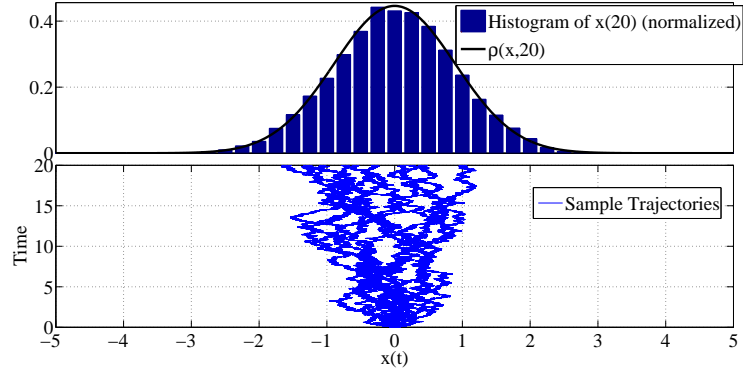
If all sample paths are started at the origin, we have $p(x, t_0) = \delta(x)$. The initial value problem (6.4.14-6.4.15) can be solved (e.g., by Fourier transforms) to yield a Gaussian distribution whose variance increases linearly with t :

$$p(x, t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{x^2}{2\sigma^2 t}\right), \quad (6.4.16)$$

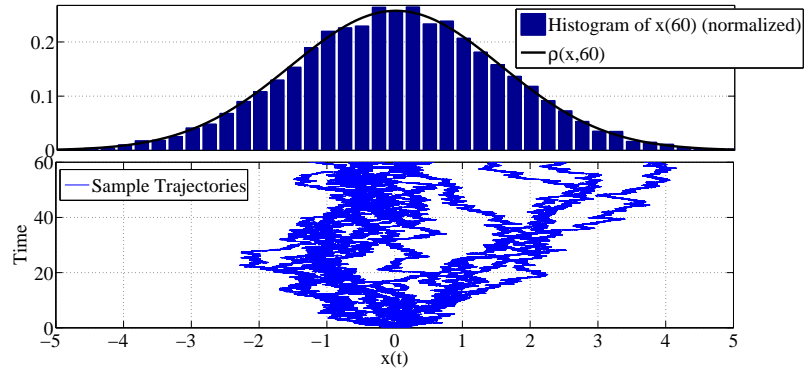
where, for simplicity, we take $t_0 = 0$. As expected, Eqn. (6.4.16) encodes the temporal signature of spreading trajectories of Brownian particles. This example provides an intuitive explanation of the diffusion term in Eqn. (6.4.12).

Fig. 87 shows sample paths (lower traces) and normalized histograms of the positions $x(T)$ generated by 10,000 simulations of the SDE (6.4.14) at times (a) $T = 20$ and (b) $T = 60$, in comparison with plots of the probability density function (6.4.16) at the same times (upper traces). Agreement between the densities $p(x, T)$ and the histograms is excellent, but note that many sample paths are necessary to obtain such good estimates, while a single solution of the Fokker-Planck equation suffices, and moreover explicitly shows how parameters in the SDE appear in the evolving densities. To further illustrate this point, consider an OU process with time-varying drift rate $A(t)$:

$$dx = [\lambda x + A(t)] dt + \sigma dW, \quad x(0) = x_0. \quad (6.4.17)$$



(a)



(b)

Figure 87: Brownian Motion. Lower panels in each block show sample paths of (6.4.14) with horizontal axis $x(t)$ and vertical axis t . Upper panels plot solutions (6.4.16) of the Fokker-Planck equation (black traces) along with normalized histograms of sample paths evaluated at (a) $T = 20$ and (b) $T = 60$. Sample paths were computed by the Euler-Maruyama method (§6.4.4).

(For $\lambda = 0$ this becomes a pure drift-diffusion (DD) process, although some authors reserve that term for SDEs with A constant.) In this case the Fokker-Planck equation is

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x} [(\lambda x + A(t))p] + \frac{\sigma^2}{2} \frac{\partial^2 p}{\partial x^2}, \quad (6.4.18)$$

The term involving the first spatial derivative now appears due to the nonzero drift A and λx , which exert a deterministic force. In the special case $\sigma = 0$, $\lambda = 0$ and A constant, we have convection at speed A :

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x} [A p] = -A \frac{\partial p}{\partial x} \Rightarrow p(x, t) = p(x - At, 0). \quad (6.4.19)$$

This first order PDE provides the simplest example of traveling waves: the initial data $p(x, 0)$ is propagated at speed A without change of shape (cf. §3.6.1, and remember manipulations with the traveling coordinate $\zeta = x - At$ and the chain rule). This provides intuition for the drift term in Eqn. (6.4.12).

In general, for a Gaussian distribution of initial data

$$p(x, 0) = \frac{1}{\sqrt{2\pi\nu_0}} \exp \left[-\frac{(x - \mu_0)^2}{2\nu_0} \right], \quad (6.4.20)$$

the solution of (6.4.18) is:

$$p(x, t) = \frac{1}{\sqrt{2\pi\nu(t)}} \exp \left[-\frac{(x - \mu(t))^2}{2\nu(t)} \right], \text{ where} \quad (6.4.21)$$

$$\mu(t) = \mu_0 e^{\lambda t} + \int_0^t e^{\lambda(t-s)} A(s) ds \quad \text{and} \quad \nu(t) = \nu_0 e^{2\lambda t} + \frac{c^2}{2\lambda} (e^{2\lambda t} - 1). \quad (6.4.22)$$

In the case of stable OU processes ($\lambda < 0$), the variance $\nu(t)$ converges to $c^2/|2\lambda|$, while for unstable processes ($\lambda > 0$) $\nu(t)$ grows exponentially fast. For pure drift diffusion processes ($\lambda = 0$), the mean and variance formulae simplify to:

$$\mu(t) = \mu_0 + \int_0^t A(s) ds \quad \text{and} \quad \nu(t) = \nu_0 + c^2 t, \quad (6.4.23)$$

giving linear growth, as in Brownian motion. The time-varying drift term $A(t)$ affects only the center of mass or average of solutions of (6.4.1), which evolve according to the deterministic ODE

$$\dot{\mu} = \lambda \mu + A(t), \quad \mu(0) = \mu_0. \quad (6.4.24)$$

Exercise 48. Verify that Eqns. (6.4.21)-(6.4.22) and (6.4.23) satisfy the forward Kolmogorov equation (6.4.18) with initial condition (6.4.20).

Exercise 49. Use the results above to verify the claim made following Eqn. (6.1.5): that solutions of the stable OU process (6.1.5a) approach a Gaussian distribution with mean and variance as stated. Describe what happens if $\lambda = w - k > 0$ in (6.1.5b).

The Kolmogorov equation (6.4.12) generalises to SDEs with multi-dimensional state spaces in a natural way: the partial derivatives $\partial/\partial x$ and $\partial^2/\partial x^2$ becoming divergence and Laplacian operators respectively [71]. To illustrate this, and connect to the LAM of §6.1.2, in Fig. 88 we

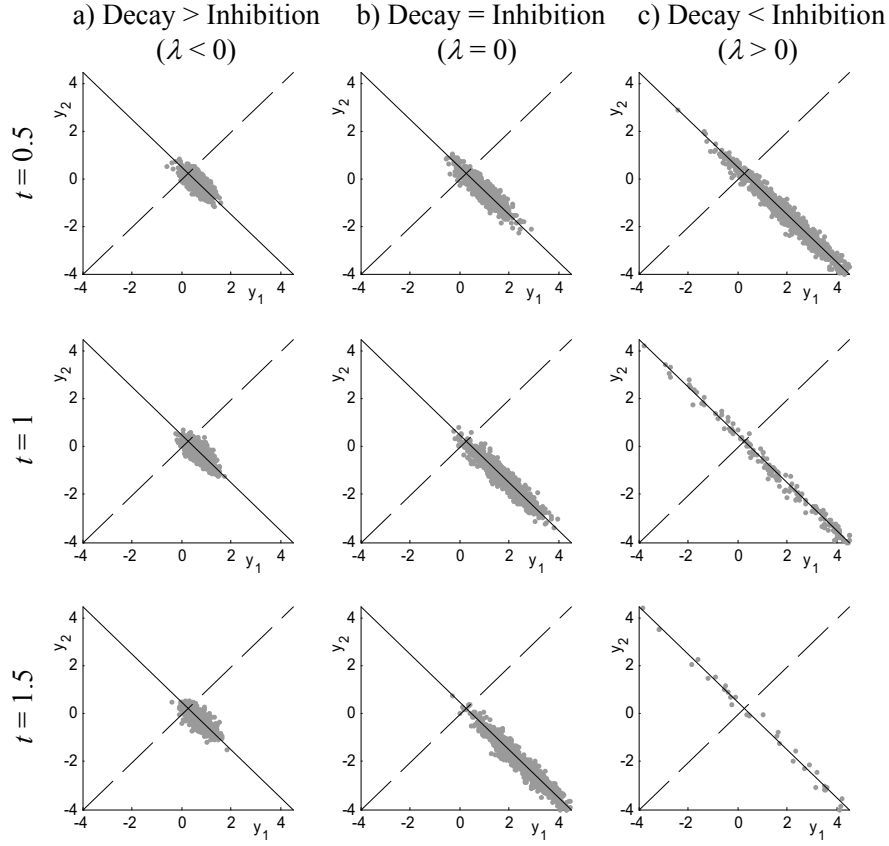


Figure 88: Distributions of 1000 sample paths for the linearised leaky accumulator model (6.1.3) at early (0.5 sec, top row), middle (1.0 sec, middle row) and later (1.5 sec, bottom row) times with $k = 11.5 > w = 8.5$ (left column), $k = w = 10.0$ (middle column) and $k = 8.5 < w = 11.5$ (right column). Attracting line shown thin solid. Remaining parameters values are $s_1 = 3 + 2.19\sqrt{2}$, $s_2 = 3$, $c = 1$. In lower right most solutions have left the region shown. From [17, Fig. 11]; sample paths were computed by the Euler-Maruyama method (§6.4.4). **NOTE: Interpret** $y_j \mapsto x_j$.

show numerically-simulated sample paths for Eqn. (6.1.3) at three times after starting with a delta function distribution on the attracting line at $y_1 = (s_1 + s_2)/[\sqrt{2}(w + k)]$, $y_2 = 0$ in the eigenvector coordinates. Thus, the density projected onto the attracting line is effectively that for the scalar OU-DD process (6.1.5b) and the density projected onto the dashed diagonal is that for the stable OU process (6.1.5a). The left and right hand columns shows cases of stable ($\lambda < 0$) and unstable ($\lambda > 0$) OU processes respectively in the y_2 direction, and the central column is the pure DD process (a balanced LAM).

6.4.3 First passage problems and the backward Kolmogorov equation

As we shall show in §6.5.2, the expressions (6.4.21-6.4.22) allow us to predict accuracy as a function of time in 2AFC tasks run under a hard limit of the cued or deadlined protocol. Free responses, on the other hand, are usually modeled as a *first passage problem*. Indeed, this is the situation that emerges in the continuum limit of SPRT derived in §6.3.2: the decision is made when the solution of the SDE (6.3.21) first reaches one of the two thresholds $+z$ or $-z$.

We outline the first passage problem for the general nonlinear SDE

$$dx = f(x) dt + g(x) dW , \quad (6.4.25)$$

in which the functions $f(x)$ and $g(x)$ depend on time only via the state variable $x(t)$, and passage of $x(t)$ through $x = a$ (resp., $x = b$), with $b > a$ and $x(0) \in (a, b)$, corresponds to an incorrect (resp., correct) decision (to fix ideas, think of the case $f(x) > 0$). First passage probabilities $\Pi(x, t)$ are governed by the *backward Kolmogorov* or *Fokker-Planck equation*:

$$\frac{\partial \Pi}{\partial t} = f(x) \frac{\partial \Pi}{\partial x} + \frac{g^2(x)}{2} \frac{\partial^2 \Pi}{\partial x^2} , \quad (6.4.26)$$

To distinguish from the forward evolution problem (6.4.12) we use the variable $\Pi(x, t)$ in place of $p(x, t)$. Note that, in (6.4.26), the functions $f(x)$ and $g(x)$ are now *outside* the partial x -derivatives. Derivations of Eqn. (6.4.26) and of the ODEs that follow are more complicated than for the forward equation (6.4.12) and we do not sketch them here; they can be found in [71].

Specifically, for the homogeneous problem (6.4.25) with time-independent coefficients, the probability $\Pi_a(x_0)$ of first passage being through $x = a$, given a starting point $x_0 \in (a, b)$, is found from the following boundary value problem:

$$f(x_0)\Pi'_a(x_0) + \frac{g^2(x_0)}{2}\Pi''_a(x_0) = 0, \quad \Pi_a(a) = 1, \quad \Pi_a(b) = 0. \quad (6.4.27)$$

(see [71, Eqn. (5.2.186)]). Note that the requirement of time-independent coefficients implies that the variable drift case $A(t)$ in (6.4.1) is *not* covered by this theory.

Multiplying the linear ODE (6.4.27) by the integrating factor

$$\psi(x) = \exp \left(\int_a^x \frac{2f(y)}{g^2(y)} dy \right) , \quad (6.4.28)$$

we may integrate once with respect to x_0 to obtain:

$$\frac{d}{dx_0} [\Pi'_a(x_0)\psi(x_0)] = 0 \Rightarrow \Pi'_a(x_0)\psi(x_0) - \Pi'_a(a)\psi(a) = c_0 \Rightarrow \Pi'_a(x_0) = \frac{c_0}{\psi(x_0)}, \quad (6.4.29)$$

since $\psi(a) = 0$. Integrating once more and using the boundary conditions we find:

$$\Pi_a(x_0) = c_0 \int_a^{x_0} \frac{dy}{\psi(y)} + c_1 \Rightarrow \Pi_a(a) = c_1 = 1, \quad \Pi_a(b) = c_0 \int_a^b \frac{dy}{\psi(y)} + c_1 = 0, \quad (6.4.30)$$

which determines the two constants of integration $c_0 = -1/\int_a^b \frac{dy}{\psi(y)}$ and $c_1 = 1$. Hence the error rate is

$$\text{ER} = \Pi_a(x_0) = 1 - \frac{F_a^{x_0}}{F_a^b} = \frac{F_a^b - F_a^{x_0}}{F_a^b} = \frac{F_{x_0}^b}{F_a^b}, \quad \text{where } F_{x_1}^{x_2} = \int_{x_1}^{x_2} \frac{dy}{\psi(y)}. \quad (6.4.31)$$

(Note that equations (5.2.189) and (5.2.190) of [71] are incorrect.) The accuracy or probability that a correct decision is made is therefore

$$1 - \text{ER} = 1 - \frac{F_{x_0}^b}{F_a^b} = \frac{F_a^b - F_{x_0}^b}{F_a^b} = \frac{F_a^{x_0}}{F_a^b}. \quad (6.4.32)$$

The mean first passage time $T(x_0)$ for starting point x_0 and passage through $x = a$ or $x = b$, is found from a second boundary value problem:

$$f(x_0)T'(x_0) + \frac{g^2(x_0)}{2}T''(x_0) = -1, \quad T(a) = T(b) = 0, \quad (6.4.33)$$

(see [71, Eqn. (5.2.154)]), which also follows from the backward Kolmogorov equation and may be solved much as above. (A simple derivation of Eqn. (6.4.33) is given in [61, §10.1.6.1].) Letting

$$h(y) = \frac{2}{\psi(y)} \int_a^y \frac{\psi(s)}{g^2(s)} ds, \quad G_{x_1}^{x_2} = \int_{x_1}^{x_2} h(y) dy, \quad (6.4.34)$$

we find that the mean first passage time is [71, Eqn. (5.1.158)]:

$$T(x_0) = \frac{F_a^{x_0} G_{x_0}^b - F_{x_0}^b G_a^{x_0}}{F_a^b}. \quad (6.4.35)$$

In the following, we shall interpret this as the *mean decision time*, i.e., $T(x_0) = \langle \text{DT} \rangle$, since the DD process models only the evidence accumulation part of the cognitive process. The reaction time (RT) also includes sensory transduction and motor response components, which are usually modelled by an additive *non-decision time* T_{nd} , so that $\langle \text{RT} \rangle = \langle \text{DT} \rangle + \langle T_{nd} \rangle$ [174, 203].

Although Eqns. (6.4.31) and (6.4.35) give explicit expressions for ER and $\langle \text{DT} \rangle$, the integrals $F_{x_1}^{x_2}$ and $G_{x_1}^{x_2}$ can only be evaluated in closed form for simple drift and diffusion coefficients $f(x), g(x)$. Even for the OU process (6.4.17) with constant A complicated special functions are required [17, Appendix]. Fortunately, for the pure DDM (6.3.21), simple expressions emerge: see §6.5.1 below).

6.4.4 Numerical integration: the Euler-Maruyama method

As we know, few nonlinear equations can be solved explicitly, and we frequently resort to numerical integration to obtain approximate solutions. We end this section by describing the simplest integration method for stochastic ODEs, which generalizes the forward Euler method of §2.3.1.

Taking steps of fixed size Δt and letting $t_n = t_0 + n\Delta t$ and $x_n = x(t_0 + n\Delta t)$, the SDE (6.4.2) is discretized as follows:

$$\begin{aligned} x_{n+1} &= x_n + f(x_n, t_n) \Delta t + g(x_n, t_n) dW(t_n) \quad \text{or} \\ x_{n+1} &= x_n + f(x_n, t_n) \Delta t + g(x_n, t_n) \sqrt{\Delta t} N(0, 1). \end{aligned} \quad (6.4.36)$$

Here we again use the diffusion property (6.4.4) of the Wiener process in drawing samples of size $\mathcal{O}(\sqrt{\Delta t})$ from a normal distribution. Note that the function $g(x, t)$ is evaluated at the start (t_n) of each time step, implying that the random increments are summed as in the Itô integral (6.4.3). For more discussion and examples of Matlab codes that execute the Euler-Maruyama method, see [99], and for extensive treatment of numerical methods for SDEs [125]. **CITE an introductory text, if there's a good one!**

In §2.3.1 we proved that solutions of the forward Euler method, defined on a finite interval $[t_0, t_0 + T]$ for the deterministic ODE $\dot{x} = f(x, t)$, converge on the true solution as $\Delta t \rightarrow 0$, with a global accumulated error of $\mathcal{O}(\Delta t)$. More precisely and in the present context, letting $x(t)$ be the true solution of $\dot{x} = f(x, t)$ and $\{x_n\}_{n=0}^N$ be the discrete sequence generated by Eqn (6.4.36) with $g(x, t) \equiv 0$, we proved first order convergence in the sense that

$$|x_N - x(t_0 + T)| \leq \mathcal{K} \Delta t, \quad (6.4.37)$$

where the prefactor \mathcal{K} depends upon the properties of $f(x, t)$ and the elapsed time T , cf. Eqn. (2.3.17).

Since sample paths $x(t)$ and x_n of both the SDE (6.4.2) and its discrete approximation (6.4.36) are random variables, we must use expected values to gauge the error in the present case. If

$$\mathbb{E}[|x_n - x(t_0 + n\Delta t)|] \leq \mathcal{K} \Delta t^\gamma \quad (6.4.38)$$

for all $n \in [0, T/\Delta t]$, the numerical method is said to have *strong order of convergence* γ . Here \mathcal{K} depends upon T and properties of $f(x, t)$, $g(x, t)$, but for appropriate functions it can be shown that the Euler-Maruyama method satisfies inequality (6.4.38) with $\gamma = 1/2$. Recalling the fact that increments $dW \sim \sqrt{\Delta t}$, this is not surprising.

As noted in [99], Eqn. (6.4.38) bounds the rate at which the *mean of the errors* (averages over many sample paths) decreases as $\Delta t \rightarrow 0$. One can also seek a bound for the rate of decrease in the *error of the means*:

$$|\mathbb{E}[x_n] - \mathbb{E}[x(t_0 + n\Delta t)]| \leq \mathcal{K} \Delta t^\gamma. \quad (6.4.39)$$

If inequality (6.4.39) holds for all $n \in [0, T/\Delta t]$, the method has *weak order of convergence* γ . The fact that $\mathbb{E}[dW] = 0$ implies that the noise terms average to zero in taking expectations of the integral equation (6.4.2) and its discretized analogue (6.4.36). In other words, the expected values of sample paths evolve according to the deterministic parts of the SDE (6.4.1) and the discrete map (6.4.36); cf. Eqn. (6.4.24). The weak order of convergence for the Euler-Maruyama method is therefore $\gamma = 1$. Numerical investigations of strong and weak convergence are described and displayed in [99].

6.5 A return to two-alternative forced-choice tasks

As we noted in §6.1, the 2AFC can be administered under a *free response protocol*, in which subjects respond in their own time, deadlines may be imposed (to encourage rapid responses), or the subject may be instructed to respond when a cue is presented, possibly some time after the stimulus display is removed, as in a working memory task. A simplified limit of cued responses is obtained by assuming that the stimulus remains visible until the response cue is presented, at which point the subject is supposed to respond without further delay. We call this the *interrogation protocol*. The DDM can model both these situations, as we now show.

6.5.1 The free response protocol

The only case in which simple formulae for error rates and mean decision times is the DDM with constant drift:

$$dy = A dt + c dW, \quad y(0) = y_0, \quad (6.5.1)$$

and to further simplify we shall also assume symmetric thresholds $y = \pm z$. Without loss of generality, we take $A > 0$, so that passage through $y = +z$ indicates a correct response, and through $-z$ an incorrect response. Letting $y = Ax$ and defining new parameters

$$\alpha = \frac{z}{A} > 0, \quad \beta = \left(\frac{A}{c}\right)^2 > 0, \quad (6.5.2)$$

equation (6.5.1) becomes

$$dx = dt + \frac{1}{\sqrt{\beta}} dW, \quad x(0) = \frac{y_0}{A} = x_0, \quad (6.5.3)$$

with thresholds $x = \pm\alpha$. The rescaling reveals that the problem only depends on *two* parameters: the threshold-to-drift ratio α and the signal-to-noise ratio (SNR) β (Note that α is the time taken to reach threshold in the absence of noise: $\text{SNR} = \infty$).

Eqn. (6.5.3) is a special case of (6.4.25) with $g(x) = 1$ and $D(x) = 1/\beta$. Hence, from Eqns. (6.4.28), (6.4.31) and (6.4.34) we can calculate that

$$\psi(y) = e^{2\beta(y+\alpha)}, \quad h(y) = 1 - e^{-2\beta(y+\alpha)},$$

and

$$F_a^{x_0} = \frac{1}{2\beta} \left(1 - e^{-2(\alpha+x_0)\beta}\right), \quad F_{x_0}^b = \frac{e^{-2\alpha\beta}}{2\beta} \left(e^{-2x_0\beta} - e^{-2\alpha\beta}\right), \quad F_a^b = \frac{1}{2\beta} \left(1 - e^{-4\alpha\beta}\right),$$

$$G_a^{x_0} = \alpha + x_0 + \frac{1}{2\beta} \left(e^{-2(\alpha+x_0)\beta} - 1\right), \quad G_{x_0}^b = \alpha - x_0 + \frac{1}{2\beta} e^{-2\alpha\beta} \left(e^{-2\alpha\beta} - e^{-2x_0\beta}\right).$$

Thus, via (6.4.31) and (6.4.35) we have:

$$\text{ER} = \frac{1}{1 + e^{2\alpha\beta}} - \left\{ \frac{1 - e^{-2x_0\beta}}{e^{2\alpha\beta} - e^{-2\alpha\beta}} \right\}, \quad (6.5.4)$$

$$\langle \text{DT} \rangle = \alpha \tanh(\alpha\beta) + \left\{ \frac{2\alpha(1 - e^{-2x_0\beta})}{e^{2\alpha\beta} - e^{-2\alpha\beta}} - x_0 \right\}, \quad (6.5.5)$$

Note that the error rate (6.5.4) may be made as small as we wish for a given drift and noise variance, by picking $\alpha = z/A$ sufficiently high. Also, biased initial data with $x_0 > 0$ reduces the ER and mean DT (since we have assumed that $A > 0$ and the correct threshold $+z > 0$ is closer), while if $x_0 < 0$ the ER increases. If the two stimuli are delivered with equal probability and we require equal accuracy for responses to both, the optimal procedure is to start each trial with unbiased initial data: $x_0 = 0$. (Recall that the analogous SPRT procedure starts with $I_0 = \log(1) = 0$, cf. Eqn. (6.3.10).) In this case Eqns (6.5.4-6.5.5) reduce to the simple expressions

$$\text{ER} = \frac{1}{1 + e^{2\alpha\beta}} \quad \text{and} \quad \langle \text{DT} \rangle = \alpha \tanh(\alpha\beta) = \alpha \left(\frac{e^{2\alpha\beta} - 1}{e^{2\alpha\beta} + 1} \right). \quad (6.5.6)$$

6.5.2 The interrogation protocol

Cued responses can be modeled by assuming that sample paths of the drift-diffusion process evolve freely until interrogation, at which instant we interpret the probability of responses in favor of hypotheses H_0 (resp., H_1) by asking if a given sample path lies above or below $y = 0$. This is the continuum analog of the Neyman-Pearson test. Specifically, we evaluate the integrals of the probability distribution of solutions of the forward Kolmogorov equation between $-\infty$ and 0 and 0 and $+\infty$ respectively, to evaluate the expected %-correct and error rates. In the interval before interrogation a sample path may cross and recross $y = 0$ many times.

Assuming again that $y > 0$ represents the correct choice, the probabilities of correct and incorrect choices at time t are therefore:

$$\mathcal{P}(\text{correct}) = 1 - \text{ER} = \int_0^\infty p(y, t) dy \quad \text{and} \quad \mathcal{P}(\text{incorrect}) = \text{ER} = \int_{-\infty}^0 p(y, t) dy. \quad (6.5.7)$$

Using the change of variables

$$u = \frac{y - \mu}{\sqrt{2\nu}},$$

and the expression (6.4.21) derived above for $p(y, t)$, the first of these integrals becomes

$$\mathcal{P}(\text{correct}) = \int_{-\frac{\mu}{\sqrt{2\nu}}}^\infty \frac{1}{\sqrt{\pi}} e^{-u^2} du = \frac{1}{2} \left[1 + \text{erf} \left(\frac{\mu}{\sqrt{2\nu}} \right) \right], \quad (6.5.8)$$

and consequently we also have

$$\mathcal{P}(\text{incorrect}) = \text{ER} = \frac{1}{2} \left[1 - \text{erf} \left(\frac{\mu}{\sqrt{2\nu}} \right) \right]. \quad (6.5.9)$$

Here erf denotes the error function integral ($\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$; $\text{erf}(\infty) = 1$), and the appropriate expressions for the evolving mean $\mu(t)$ and variance $\nu(t)$ are found from (6.4.22) or (6.4.23). The expression $\mathcal{P}(\text{correct})$ of (6.5.7) is therefore a function of the time t at which the response is given. In psychology this cumulative distribution is called the *psychometric function*.

6.6 Optimal performance on a free response task: analysis and predictions

In §6.1 we referred to the *speed-accuracy tradeoff*. In many situations there are competing pressures for accuracy, and speed: one would like to finish the test *and* get as many questions right as possible. How long should one spend on each question? The constrained 2AFC situation allows us to derive the optimal compromise using little more than freshman calculus.

We have seen that the neurally-inspired (albeit not rigorously-derived) leaky accumulator model reduces to a DDM in the limit of balanced, and sufficiently large, leak and inhibition. It is therefore plausible that human and animal brains may be able to approximate DDM decision dynamics. Moreover the DDM is the optimal decision maker in the sense of Theorem 5: faced with noisy incoming data it delivers a decision of guaranteed accuracy in the shortest possible time. It remains to balance accuracy and speed, which, as we shall see, involves selecting the appropriate threshold depending on task conditions.

Suppose that each correct response is rewarded with a fixed sum of money (for humans) or a drop of juice (for thirsty monkeys), both assumed to be items of fixed value, and that each block of trials runs for a fixed duration. If the subject wishes to maximise his or her total reward, then the average *reward rate* given by the following function must be maximised:

$$\text{RR} = \frac{1 - \text{ER}}{\langle \text{DT} \rangle + D + D_p \text{ER}}. \quad (6.6.1)$$

The numerator is the probability of being correct on each trial, and the denominator is the average time that elapses between each response. Here D denotes the response to stimulus interval (RSI) and D_p denotes an additional penalty delay that may be imposed following errors. Both D and D_p are set by the experimenter, and the difficulty of the task may also be adjusted, e.g. via coherence in the moving dots stimulus. To maintain a stationary environment, the delays D , D_p and the difficulty must be kept constant during each block of trials. If we assume that the latter implies a constant SNR $\sqrt{\beta}$, then we may derive optimal performance as follows¹⁷.

We first substitute the expressions (6.5.6) for ER and $\langle \text{DT} \rangle$ into (6.6.1). A short calculation yields the expression:

$$\text{RR} = \left[\alpha + D + (D + D_p - \alpha) e^{-2\alpha\beta} \right]^{-1}, \quad \text{where } \alpha = \frac{z}{A} \text{ and } \beta = \left(\frac{A}{c} \right)^2 \quad (6.6.2)$$

from Eqn. (6.5.2). Given fixed values of D , D_p and SNR ($\Rightarrow \beta = \text{const.}$), we can now maximize RR, or rather minimize $1/\text{RR}$, with respect to the threshold-to-drift ratio α :

$$\frac{d}{d\alpha} \left(\frac{1}{\text{RR}} \right) = 1 - e^{-2\alpha\beta} - 2\beta(D + D_p - \alpha) e^{-2\alpha\beta} = 0 \quad \Rightarrow \quad e^{2\alpha\beta} - 1 = 2\beta(D + D_p - \alpha). \quad (6.6.3)$$

Eqn. (6.6.3) has a unique solution $\alpha = \alpha(\beta, D + D_p)$ that depends only on SNR and the *sum* of the RSI and penalty delay, and this critical point is a minimum of $1/\text{RR}$, and hence a maximum of RR. See Fig. 89.

¹⁷Note that many experiments, especially those using primates, are run with coherences and viewing times drawn from random distributions, so that the local temporal environments are *not* stationary in the sense required here.

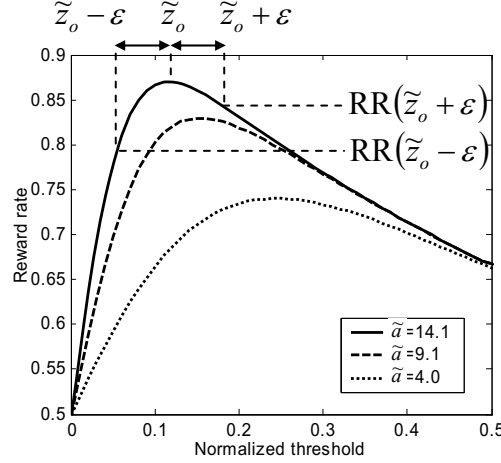


Figure 89: Reward rate as a function of rescaled threshold α for three values of SNR: $\beta = 4.0$ (dotted), $\beta = 9.1$ (dashed) and $\beta = 14.1$ (solid). Perturbations ϵ from optimal threshold that maximizes RR are described in text at the end of this section. From [17, Fig. 15].

Exercise 50. Check that the critical point of the function $1/\text{RR}$ found above is a minimum, and show that it is the only critical point of this function. Plot graphs of the reward rate RR as a function of α for values of $\beta = 10, 5, 1$ and $D = 0.5, 1.0, 2$ (set $D_p = 0$).

We may go further, solving for α and β in terms of $\langle \text{DT} \rangle$ and ER from (6.5.6) to obtain

$$\alpha = \frac{\langle \text{DT} \rangle}{1 - 2\text{ER}}, \quad \beta = \frac{1 - 2\text{ER}}{2\langle \text{DT} \rangle} \log \left(\frac{1 - \text{ER}}{\text{ER}} \right). \quad (6.6.4)$$

Substituting Eqns. (6.6.4) into (6.6.3) yields the *speed-accuracy tradeoff* that follows from maximizing RR:

$$\frac{\langle \text{DT} \rangle}{D_{\text{total}}} = \left[\frac{1}{\text{ER} \log \left(\frac{1 - \text{ER}}{\text{ER}} \right)} + \frac{1}{1 - 2\text{ER}} \right]^{-1}, \quad (6.6.5)$$

where $D_{\text{total}} = D + D_p$. This *optimal performance curve* (OPC) uniquely relates the mean normalised decision time $\langle \text{DT} \rangle / D_{\text{total}}$ to ER . Note that only experimental observables appear in Eqn (6.6.5): there are no other parameters, and in particular, no free parameters. Hence data collected for different subjects (who may exhibit differing SNRs, even when viewing the same stimuli), and for differing RSIs and penalty delays, can be pooled and compared with the theory.

The form of the OPC, shown in Fig. 90(a), may be intuitively understood by observing that very noisy stimuli ($\beta \approx 0$) contain little information, so that given a priori knowledge that they are equally likely, it is optimal to choose at random without examining them, giving $\text{ER} = 0.5$ and $\langle \text{DT} \rangle = 0$ (note the points for $\beta = 0.1$ in Fig. 90(a)). At the opposite limit $\beta \rightarrow \infty$, noise-free stimuli are so easy to discriminate that both $\langle \text{DT} \rangle$ and ER approach zero (note the points for $\beta = 100$ in Fig. 90(a)). This limit yields the highest RRs, but when SNR is finite, due to poor stimulus discriminability or a subject's inability to clearly detect a signal, making immediate decisions is

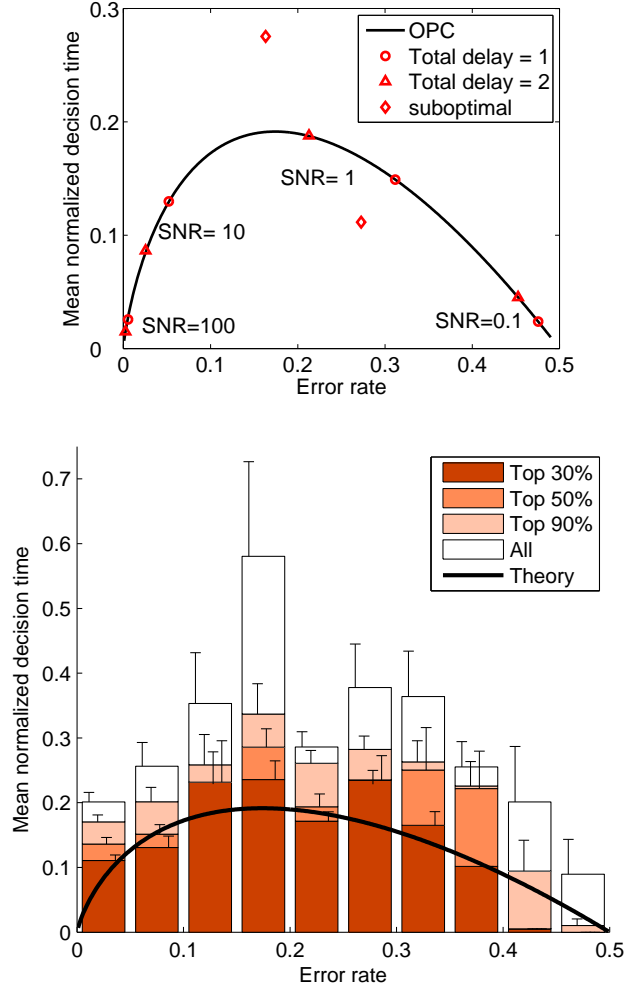


Figure 90: Top: Optimal performance curve (OPC) of Eq. (6.6.5) relating mean normalized decision time to error-rate across varying task conditions. Triangles and circles denote performances under conditions specified in box. Moving left, RRs increase with SNR from 0.51 to 0.60, 0.84 and 0.97 with $D_{tot} = 1$, and from 0.26 to 0.33, 0.45 and 0.49 with $D_{tot} = 2$. Suboptimal performance results from thresholds 25% above and below optimal threshold with SNR=1 and $D_{tot} = 2$ (diamonds); both $\approx 1.3\%$ below maximum RR. Bottom: OPC (black curve) compared with data from 80 human subjects (histograms), sorted according to total rewards accrued over multiple blocks of trials with two difficulty levels and $D = 0.5, 1.0, 2.0$ s and $D_p = 0, 1.5$ s. Open bars: all subjects; light bars: lowest 10% excluded; medium bars: lowest 50% excluded; dark bars: lowest 70% excluded. Vertical line segments indicate standard errors. From [237, Fig. 1].

not optimal. Instead, it is advantageous to accumulate the noisy evidence for just long enough to make the best possible choice (see the points for $\beta = 1$ and $\beta = 10$ in Fig. 90(a)). Non-optimal thresholds degrade performance, as illustrated by the diamonds in Fig. 90(a), which result from setting α 25% above and below optimal for $\beta = 1$ and $D_{tot} = 2$: in both cases the RR degrades by $\approx 1.3\%$. See [17, 237] for further details.

Fig. 90(b) shows a histogram of behavioral data compiled from human subjects, indicating that those who score in the top 30% overall on a series of tests with differing delays and SNRs follow the optimal curve remarkably closely [109, 18]. More detailed data analysis in [18] reveals that, in each block of trials for which stimulus recognition difficulty (\sim SNR) and RSI are held constant, these subjects rapidly adjust their thresholds to achieve this. (A model for threshold adjustment is described in [201].)

However, other subjects, and especially the lowest-scoring 10%, display suboptimal behavior, with significantly longer decision times and correspondingly lower ERs. Previous studies have shown that humans often favor accuracy over reward [157], and alternative objective functions have been proposed to account for this. For example, one can define a modified reward rate, weighted toward accuracy by additionally penalizing errors, as suggested by the proposal that humans exhibit a competition between reward and accuracy (COBRA) [146, 19]:

$$RA = RR - \frac{q}{D_{total}} ER; \quad (6.6.6)$$

here the factor q specifies the additional weight placed on accuracy, and the characteristic time D_{total} is included in the second factor, so that the units of both terms in RA are consistent. Maximizing RA as above we obtain a family of OPCs parameterized by q :

$$\frac{DT}{D_{total}} = \frac{\mathcal{E} - 2q - \sqrt{\mathcal{E}^2 - 4q(\mathcal{E} + 1)}}{2q}, \quad (6.6.7)$$

where

$$\mathcal{E} = \left[\frac{1}{ER \log\left(\frac{1-ER}{ER}\right)} + \frac{1}{1-2ER} \right]. \quad (6.6.8)$$

If rewards are monetary, one can also postulate a situation in which errors are rewarded (albeit less lavishly than correct responses), or penalized by subtraction from previous winnings:

$$RR_m = \frac{(1 - ER) - qER}{DT + D_{total}} \quad (6.6.9)$$

This leads to the following family of OPCs:

$$\frac{DT}{D_{total}} = (1 + q) \left[\frac{\frac{1}{ER} - \frac{q}{1-ER}}{\log\left(\frac{1-ER}{ER}\right)} + \frac{1 - q}{1 - 2ER} \right]^{-1}. \quad (6.6.10)$$

Eqns. (6.6.7-6.6.8) and Eqn. (6.6.10) both reduce to (6.6.5) for $q = 0$, as expected. Fig. 91 shows an example of the second family (6.6.10). Eqn. (6.6.7) gives a similar family, but the maxima remain at the same ER (≈ 0.18) as q increases, rather than moving rightwards as in Fig. 91 [17, 18].

Both these functions involve a weight parameter q , which will typically be subject-dependent, since different people may place a greater or lesser weight on accuracy, even if they understand that a specific balance is implied, as in Eqn. (6.6.9). Values of q should therefore be fitted to individuals or subgroups of subjects, and the theory becomes descriptive rather than prescriptive. However, Fig. 91 shows that an average weight ($q = 0.62$) may be assigned to the entire group to yield an improved fit over that to the RR in Fig 90. See [18] for further details and fits of these curves to behavioral data.

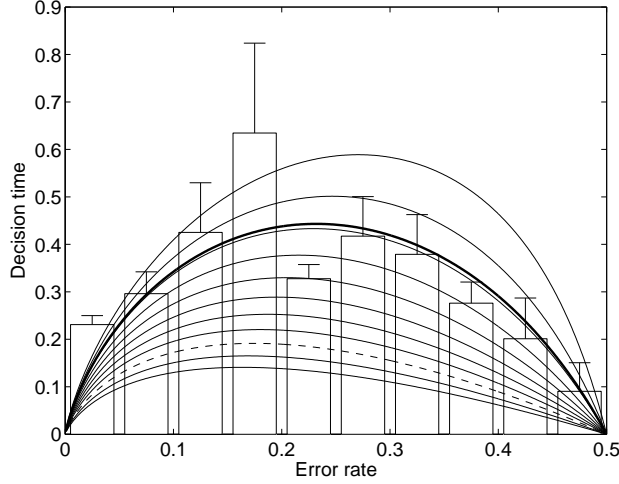


Figure 91: Optimal performance curves of (6.6.10) for the modified reward rate function RR_m of (6.6.9) with q varied in steps of 0.1 between -0.2 (lowest curve) and 0.8 (highest curve). The dashed curve corresponds to $q = 0$ (Eqn. (6.6.5)) and the bold solid curve to $q = 0.62$: the best fit to *all* the subjects in the study (white bars). Error bars indicate standard error. From [109, Fig. 2].

One can also observe that the typical shape of the RR vs. threshold curve (see Fig. 89 above) is asymmetric around its maximum, falling more steeply toward lower thresholds and less steeply toward higher thresholds. The loss of reward is therefore greater for underestimates of optimal thresholds than for overestimates, assuming misestimation is symmetric ($RR(\tilde{z}_0 - \epsilon) < RR(\tilde{z}_0 + \epsilon)$) [17]. This may partially explain the tendency of some subjects to pick higher thresholds and consequently emphasise accuracy. Indeed, a re-analysis of the data of [18], suggested by information gap theory [14], shows that such subjects may be allowing for their uncertainty in assessing how much time has elapsed between responses [237]. This was modeled by allowing the RSIs to belong to a closed interval \mathcal{I}_D of finite length and maximizing the worst RR that occurs for $D \in \mathcal{I}_D$.

A recent study that allowed more training sessions than [18] found that, while most subjects started the experiment with a bias toward accuracy, their performance approached the OPC over multiple sessions as they abandoned this bias, *and that* the remaining individual deviations from the OPC were correlated with subjects' coefficients of variation in an interval timing task [11]. However, for the range of difficulties used, it was also found that a constant threshold strategy fitted the “well-trained” data almost as well as the OPC.

Finally, experiments featuring stimuli of unequal probability and unequal rewards can also be analyzed in terms of reward rate metrics, and (again with sufficient training), both humans on a free response task [200], and monkeys in an interrogation task [63, 181] have been found to exhibit near optimal performances.

References

- [1] L.F. Abbott and E.S. Chance. Drivers and modulators from push-pull and balanced synaptic input. *Prog. Brain Res.*, 149:147–155, 2005.
- [2] L.F. Abbott and T. Kepler. Model neurons: From Hodgkin-Huxley to Hopfield. In L. Garrido, editor, *Statistical Mechanics of Neural Networks*. Springer-Verlag, Berlin, 1990.
- [3] H. Agmon-Snir, C.E. Carr, and J. Rinzel. The role of dendrites in auditory coincidence detection. *Nature*, 393:268–272, 1998.
- [4] D. J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex*, 7:237–252, 1997.
- [5] D.J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb. Cortex*, 7:237–252, 1997.
- [6] D.J. Amit and M.V. Tsodyks. Quantitative study of attractor neural network retrieving at low spike rates i: substrate-spikes, rates and neuronal gain. *Network*, 2:259–274, 1991.
- [7] J. Anderson, I. Lampl, I. Reichova, M. Carandini, and D. Ferster. Stimulus dependence of two-state fluctuations of membrane potential in cat visual cortex. *Nat. Neurosci.*, 3:617–621, 2000.
- [8] V.I. Arnold. *Geometrical Methods in the Theory of Ordinary Differential Equations*. Springer-Verlag, New York, 1983.
- [9] G. Ashida and C.E. Carr. Sound localization: Jeffress and beyond. *Curr. Opin. Neurobiol.*, 21:1–7, 2011.
- [10] A. Ayali, E. Fuchs, E. Hulata, and E. Ben Jacob. The function of inter segmental connections in determining temporal characteristics of the spinal cord rhythmic output. *Neuroscience*, 147:236–246, 2007.
- [11] F. Balci, P. Simen, R. Niyogi, A. Saxe, P. Holmes, and J.D. Cohen. Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception & Psychophysics*, 73 (2):640–657, 2011.
- [12] G.A. Barnard. Sequential tests in industrial statistics. *J. Roy. Statist. Soc. Suppl.*, 8:1–26, 1946.
- [13] T. Bayes. An essay toward solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. Lond.*, 53:370–418, 1763.
- [14] Y. Ben-Haim. *Information Gap Decision Theory: Decisions under Severe Uncertainty*. Academic Press, New York, 2006. 2nd Edition.
- [15] C.M. Bender and S.A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers*. McGraw Hill, New York, 1978.
- [16] P. Billingsley. *Probability and Measure*. Wiley, New York, 1995.

- [17] R. Bogacz, E. Brown, J. Moehlis, 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. *Psychological Review*, 113 (4):700–765, 2006.
- [18] R. Bogacz, P. Hu, P. Holmes, and J.D. Cohen. Do humans produce the speed-accuracy tradeoff that maximizes reward rate? *Quart. J. Exp. Psych.*, 63 (5):863–891, 2010.
- [19] C.J. Bohil and W.T. Maddox. On the generality of optimal versus objective classifier feedback effects on decision criterion learning in perceptual categorization. *Memory & Cognition*, 31 (2):181–198, 2003.
- [20] W.E. Boyce and R.C. DiPrima. *Elementary Differential Equations and Boundary Value Problems*. Wiley, New York, 1997.
- [21] N. Brenner, S.P. Strong, R. Koberle, W. Bialek, and R. De Ruyter van Steveninck. Synergy in a neural code. *Neural Computation*, 12:1531–1552, 2000.
- [22] R. Brette. Exact simulation of integrate-and-fire models with exponential current. *Neural Computation*, 19:2604–2609, 2007.
- [23] R. Brette, M. Rudolph, T. Carnevale, M. Hines, D. Beeman, J.M. Bower, M. Diesmann, A. Morrison, P.H. Goodman, F.C. Jr. Harris, M. Zirpe, T. Natschlager, D. Pecevski, G.B. Ermentrout, M. Djurfeldt, A. Lansner, O. Rochel, T. Vieville, E. Muller, A.P. Davison, S. El Boustani, and A. Destexhe. Simulation of networks of spiking neurons: a review of tools and strategies. *J. Comput. Neurosci.*, 23:349–398, 2007.
- [24] 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.
- [25] E. Brown, J. Gao, P. Holmes, R. Bogacz, M. Gilzenrat, and J. Cohen. Simple networks that optimize decisions. *Int. J. Bifurcation and Chaos*, 15 (3):803–826, 2005.
- [26] E. Brown, J. Moehlis, and P. Holmes. On the phase reduction and response dynamics of neural oscillator populations. *Neural Computation*, 16 (4):673–715, 2004.
- [27] 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. Comput. Neurosci.*, 17 (1):5–21, 2004.
- [28] N. Brunel and V. Hakim. Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Computation*, 11:1621–1671, 1999.
- [29] N. Brunel, V. Hakim, and M.J.E. Richardson. Firing-rate resonance in a generalized integrate-and-fire neuron with subthreshold resonance. *Phys. Rev. E*, 67:051916, 2003.
- [30] N. Brunel and M.C.W. van Rossum. Lapicque’s 1907 paper: from frongs to integrate-and-fire. *Biol. Cybern.*, 97:337–339, 2007.
- [31] N. Brunel and X.-J. Wang. Effects of neuromodulation in a cortical network model. *J. Comput. Neurosci.*, 11:63–85, 2001.
- [32] J.T. Buchanan. Neural network simulations of coupled locomotor oscillators in the lamprey spinal cord. *Biol. Cybern.*, 66:367–374, 1992.

- [33] J.T. Buchanan and S. Grillner. Newly identified glutamate interneurons and their role in locomotion in the lamprey spinal cord. *Science*, 236:312–314, 1987.
- [34] A.N. Burkitt. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. *Biol. Cybern.*, 95:1–19, 2006.
- [35] A.N. Burkitt. A review of the integrate-and-fire neuron model: II. Inhomogeneous synaptic input and network properties. *Biol. Cybern.*, 95:97–112, 2006.
- [36] J.R. Busemeyer and J.T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100 (3):432–459, 1993.
- [37] M. Camperi and X.-J. Wang. A model of visuospatial working memory in prefrontal cortex: recurrent network and cellular bistability. *J. Comput. Neurosci.*, 5:383–405, 1998.
- [38] C. Capaday and C. van Vreeswijk. Direct control of firing rate gain by dendritic shunting inhibition. *J. Integr. Neurosci.*, 5:199–222, 2006.
- [39] F.S. Chance, L.F. Abbott, and A.D. Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–782, 2002.
- [40] H.J. Chiel and R.D. Beer. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci.*, 20(12):553–557, 1997.
- [41] A.H. Cohen, P. Holmes, and R.H. Rand. The nature of coupling between segmental oscillators of the lamprey spinal generator for locomotion: a model. *J. Math Biol.*, 13:345–369, 1982.
- [42] A.H. Cohen, R.H. Rand, and P. Holmes. Systems of coupled oscillators as models of central pattern generators. In A.H. Cohen, S. Rossignol, and S. Grillner, editors, *Neural Control of Rhythmic Movements in Vertebrates*, pages 333–367. Wiley, 1988.
- [43] A.H. Cohen, S. Rossignol, and S. Grillner (Eds). *Neural Control of Rhythmic Movements in Vertebrates*. Wiley, 1988.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] J.A. Connor and C.F. Stevens. Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma. *J. Physiol.*, 213:31–53, 1971.
- [48] J.A. 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.

- [49] R. Cossart, D. Aronov, and R. Yuste. Attractor dynamics of network up states in the neo-cortex. *Nature*, 423:283–288, 2003.
- [50] P. Dayan and L.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA, 2001.
- [51] M.H. DeGroot. A conversation with George A. Barnard. *Statist. Sci.*, 3:196–212, 1988.
- [52] A. Destexhe, Z.F. Mainen, and T.J. Sejnowski. Kinetic models of synaptic transmission. In C. Koch and I. Segev, editors, *Methods in Neuronal Modeling: From Ions to Networks*, chapter 1, pages 1–25. MIT Press, Cambridge, MA, 1999. second edition.
- [53] P. Eckhoff, K.F. Wong-Lin, and P. Holmes. Optimality and robustness of a biophysical decision-making model under norepinephrine modulation. *J. Neurosci.*, 29 (13):4301–4311, 2009.
- [54] P. Eckhoff, K.F. Wong-Lin, and P. Holmes. Dimension reduction and dynamics of a spiking neuron model for decision making under neuromodulation. *SIAM J. on Applied Dynamical Systems*, 10 (1):148–188, 2011.
- [55] A.V. Egorov, B.N. Hamam, E. Franssen, M.E. Hasselmo, and A.A. Alonso. Graded persistent activity in entorhinal cortex neurons. *Nature*, 420:173–178, 2002.
- [56] G.B. Ermentrout. Type I membranes, phase resetting curves, and synchrony. *Neural Computation*, 8:979–1001, 1996.
- [57] G.B. Ermentrout. *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*. SIAM, Philadelphia, 2002.
- [58] G.B. Ermentrout and N. Kopell. Frequency plateaus in a chain of weakly coupled oscillators. *SIAM J. Math. Anal.*, 15:215–237, 1984.
- [59] G.B. Ermentrout and N. Kopell. Multiple pulse interactions and averaging in systems of coupled neural oscillators. *J. Math. Biol.*, 29:195–217, 1991.
- [60] G.B. Ermentrout and N. Kopell. Parabolic bursting in an excitable system coupled with a slow oscillation. *SIAM J. Appl. Math.*, 46:233–253, 1986.
- [61] G.B. Ermentrout and D. Terman. *Mathematical Foundations of Neuroscience*. Springer-Verlag, New York, 2010.
- [62] C.P. Fall and J. Rinzel. An intracellular Ca^{2+} subsystem as a biologically plausible source of intrinsic conditional bistability in a network model of working memory. *J. Comput. Neurosci.*, 20:97–107, 2006.
- [63] S. Feng, P. Holmes, A. Rorie, and W.T. Newsome. Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS Comput. Biol.*, 5 (2):e1000284, 2009.
- [64] R. FitzHugh. Impulses and physiological states in models of nerve membrane. *Biophys. J.*, 1:445–466, 1961.
- [65] N. Foucaud-Trocme and N. Brunel. Dynamics of the instantaneous firing rate in response to changes in input statistics. *J. Comput. Neurosci.*, 18:311–321, 2005.

- [66] N. Foucaud-Trocme, D. Hansel, C. van Vreeswijk, and N. Brunel. How spike generation mechanisms determine the neuronal response to fluctuating inputs. *J. Neurosci.*, 23:11628–11640, 2003.
- [67] E. Fransen, B. Tahvildari, A.V. Egorov, M.E. Hasselmo, and A.A. Alonso. Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron*, 49:735–746, 2006.
- [68] E. Fuchs, P. Holmes, T. Kiemel, and A. Ayali. Intersegmental coordination of cockroach locomotion: Adaptive control of centrally coupled pattern generator circuits. *Frontiers in Neural Circuits*, 4:125, 2011.
- [69] S. Fusi and M. Mattia. Collective behavior of networks with linear (VLSI) Integrate and Fire Neurons. *Neural Computation*, 11:633–652, 1999.
- [70] J. Gao and P. Holmes. On the dynamics of electrically-coupled neurons with inhibitory synapses. *J. Comp. Neurosci.*, 22:39–61, 2007.
- [71] C.W. Gardiner. *Handbook of Stochastic Methods, Second Edition*. Springer Verlag, New York, 1985.
- [72] G.L. Gerstein and B. Mandelbrot. Random walk models for the spike activity of a single neuron. *Biophys. J.*, 4:41–68, 1964.
- [73] W. Gerstner. Time structure of the activity in neural network models. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 51:738–758, 1995.
- [74] W. Gerstner and W.M. Kistler. *Spiking Neuron Models*. Cambridge University Press, Cambridge, 2002.
- [75] P.A. Getting. Comparative analysis of invertebrate central pattern generators. In A.H. Cohen, S. Rossignol, and S. Grillner, editors, *Neural Control of Rhythmic Movements in Vertebrates*, chapter 4, pages 101–128. John Wiley, New York, NY, 1988.
- [76] R.M. Ghigliazza and P. Holmes. A minimal model of a central pattern generator and motoneurons for insect locomotion. *SIAM J. on Applied Dynamical Systems*, 3 (4):671–700, 2004.
- [77] R.M. Ghigliazza and P. Holmes. Minimal models of bursting neurons: How multiple currents, conductances and timescales affect bifurcation diagrams. *SIAM J. on Applied Dynamical Systems*, 3 (4):636–670, 2004.
- [78] J. Gold and M. Shadlen. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.
- [79] J.I. Gold and M.N. Shadlen. Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Science*, 5 (1):10–16, 2001.
- [80] J.I. Gold and M.N. Shadlen. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36:299–308, 2002.
- [81] M.S. Goldman, J.H. Levine, G. Major, D.W. Tank, and H.S. Seung. Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cereb. Cortex*, 13:1185–1195, 2003.

- [82] I.J. Good. Studies in the history of probability and statistics. XXXVI. A.M. Turing’s statistical work in World War II. *Biometrika*, 66:393–396, 1979.
- [83] M.D. Greenberg. *Foundations of Applied Mathematics*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [84] S. Grillner. Bridging the gap - from ion channels to networks and behaviour. *Curr. Opin. Neurobiol.*, 9:663–669, 1999.
- [85] S. Grillner, J.T. Buchanan, and A. Lansner. Simulation of the segmental burst generating network for locomotion in lamprey. *Neurosci. Letters*, 89:31–35, 1988.
- [86] S. Grossberg. Nonlinear neural networks: principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.
- [87] B. Grothe, M. Pecka, and D. McAlpine. Mechanisms of sound localization in mammals. *Physiol. Rev.*, 90:983–1012, 2010.
- [88] J. Guckenheimer. Isochrons and phaseless sets. *J. Math. Biol.*, 1:259–273, 1975.
- [89] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.
- [90] B.S. Gutkin and G.B. Ermentrout. Dynamics of membrane excitability determine interspike interval variability: a link between spike generation mechanisms and cortical spike train statistics. *Neural Computation*, 10:1047–1065, 1998.
- [91] R. Guttman, S. Lewis, and J. Rinzel. Control of repetitive firing in squid axon membrane as a model for a neuroneoscillator. *J. Physiol. (London)*, 305:377–395, 1980.
- [92] A. Hagevik and A.D. McClellan. Coupling of spinal locomotor networks in larval lamprey revealed by receptor blockers for inhibitory amino acids: neurophysiology and computer modeling. *J. Neurophysiol.*, 72:1810–1829, 1994.
- [93] D.P. Hanes and J.D. Schall. Neural control of voluntary movement initiation. *Science*, 274:427–30, 1996.
- [94] D. Hansel, G. Mato, C. Meunier, and L. Neltner. On numerical simulations of integrate-and-fire neural networks. *Neural Computation*, 10:467–483, 1998.
- [95] N.S. Harper and D. McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430:682–686, 2004.
- [96] J. Hellgren, S. Grillner, and A. Lansner. Computer simulation of the segmental neural network generating locomotion in lamprey by using populations of network interneurons. *Biol. Cybern.*, 68:1–13, 1992.
- [97] J. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation*. Addison Wesley, Reading, MA, 1991.
- [98] A.V. Herz, T. Gollisch, C.K. Machens, and D. Jaeger. Modeling single-neuron dynamics and computations: a balance of detail and abstraction. *Science*, 314:80–85, 2006.

- [99] D.J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43:525–546, 2001.
- [100] M.W. Hirsch, S. Smale, and R.L. Devaney. *Differential Equations, Dynamical Systems and an Introduction to Chaos*. Academic Press/Elsevier, San Diego, CA, 2004.
- [101] A. Hodgkin and A. Huxley. Currents carried by sodium and potassium ions through the membrane of the giant axon of *Loligo*. *J. Physiol.*, 116:449–472, 1952.
- [102] A.L. Hodgkin and A.F. Huxley. The components of membrane conductance in the giant axon of *Loligo*. *J. Physiol.*, 116:473–496, 1952.
- [103] A.L. Hodgkin and A.F. Huxley. The dual effect of membrane potential on sodium conductance in the giant axon of *Loligo*. *J. Physiol.*, 116:497–506, 1952.
- [104] A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117:500–544, 1952.
- [105] A.L. Hodgkin, A.F. Huxley, and B. Katz. Ionic currents underlying activity in the giant axon of the squid. *Arch. Sci. physiol.*, 3:129–150, 1949.
- [106] A.L. Hodgkin, A.F. Huxley, and B. Katz. Measurement of current-voltage relations in the membrane of the the giant axon of *Loligo*. *J. Physiol.*, 116:424–448, 1952.
- [107] P. Holmes, R.J. Full, D. Koditschek, and J. Guckenheimer. The dynamics of legged locomotion: Models, analyses and challenges. *SIAM Review*, 48(2):207–304, 2006.
- [108] P. Holmes, J.L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, Cambridge, U.K., 1996.
- [109] P. Holmes, E. Shea-Brown, J. Moehlis, R. Bogacz, J. Gao, G. Aston-Jones, E. Clayton, J. Rajkowski, and J.D. Cohen. Optimal decisions: From neural spikes, through stochastic differential equations, to behavior. *IEICE Transactions on Fundamentals on Electronics, Communications and Computer Science*, E88A (10):2496–2503, 2005.
- [110] G.R. Holt and C. Koch. Shunting inhibition does not have a divisive effect on firing rates. *Neural Computation*, 9:1001–1013, 1997.
- [111] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79 (8):2554–2558, 1982.
- [112] J.J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, 81 (10):3088–3092, 1984.
- [113] F.C. Hoppensteadt and E.M. Izhikevich. *Weakly Connected Neural Networks*. Springer-Verlag, New York, 1997.
- [114] D.H. Hubel and T.N. Wiesel. Functional architecture of macaque monkey visual cortex. *Proc. Roy. Soc. Lond. B*, 198:1–59, 1977.
- [115] E.M. Izhikevich. Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15:1063–1070, 2004.

- [116] C.E. Jahr and C.F. Stevens. A quantitative description of NMDA receptor-channel kinetic behavior. *J. Neurosci.*, 10:1830–1837, 1990.
- [117] C.E. Jahr and C.F. Stevens. Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J. Neurosci.*, 10:3178–3182, 1990.
- [118] L.A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, 41:35–39, 1948.
- [119] P.E. Jercog, G. Svirskis, V.C. Kotak, D.H. Sanes, and J. Rinzel. Asymmetric excitatory synaptic dynamics underlie interaural time difference processing in the auditory system. *PLoS Comput. Biol.*, 8 (6):e1000406, 2010.
- [120] D. Johnston and S. Wu. *Foundations of Cellular Neurophysiology*. MIT Press, Cambridge, MA, 1997.
- [121] C.K.R.T. Jones. *Geometric Singular Perturbation Theory*, volume 1609 of *Lecture Notes in Mathematics*. Springer Verlag, Heidelberg, 1994. C.I.M.E. Lectures.
- [122] E.G. Jones and A. Peters. *Cerebral Cortex, Functional Properties of Cortical Cells, Vol.2*. Plenum, New York, 1984.
- [123] E.R. Kandel, J.H. Schwartz, and T.M. Jessel. *Principles of Neural Science*. McGraw-Hill, New York, 2000.
- [124] J. Keener and J. Sneyd. *Mathematical Physiology*. Springer Verlag, New York, 2009. 2nd Edition, 2 Vols.
- [125] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, Berlin, 1999.
- [126] B.W. Knight. Dynamics of encoding in a population of neurons. *J. Gen. Physiol.*, 59:734–766, 1972.
- [127] B.W. Knight. The relationship between the firing rate of a single neuron and the level of activity in a population of neurons. experimental evidence for resonant enhancement in the population response. *J. Gen. Physiol.*, 59:767–778, 1972.
- [128] C. Koch. *Biophysics of computationa - information processing in single neurons*. Oxford University Press, New York, 1999.
- [129] N. Kopell and G.B. Ermentrout. Coupled oscillators and the design of central pattern generators. *Math. Biosciences*, 90:87–109, 1988.
- [130] A.A. Koulakov, S. Taghavachari, A. Kepecs, and J.E. Lisman. Model for a robust neural integrator. *Nat. Neurosci.*, 5:775–782, 2002.
- [131] V.I. Krinsky and Yu.M. Kokoz. Reduction of the Hodgkin-Huxley system to a second order system. *Biofizika*, 18 (3):506–511, 1973.
- [132] R.P. Kukillaya and P. Holmes. A hexapedal jointed-leg model for insect locomotion in the horizontal plane. *Biol. Cybern.*, 97:379–395, 2007.

- [133] R.P. Kukillaya and P. Holmes. A model for insect locomotion in the horizontal plane: Feedforward activation of fast muscles, stability, and robustness. *J. Theoretical Biol.*, 261 (2):210–226, 2009.
- [134] R.P. Kukillaya, J. Proctor, and P. Holmes. Neuro-mechanical models for insect locomotion: Stability, maneuverability, and proprioceptive feedback. *CHAOS: An Interdisciplinary Journal of Nonlinear Science*, 19 (2):026107, 2009.
- [135] G. La Camera, A. Rauch, D. Thurbon, H.-R. Luscher, W. Senn, and S. Fusi. Multiple time scales of temporal response in pyramidal and fast spiking cortical neurons. *J. Neurophysiol.*, 96:3448–3464, 2006.
- [136] D.R.J. Laming. *Information Theory of Choice-Reaction Times*. Academic Press, New York, 1968.
- [137] L. Lapicque. Recherches quantitatives sur l’excitation électrique des nerfs traitée comme une polarisation. *J. Physiol. Pathol. Gen.*, 9:620–635, 1907.
- [138] P. Latham and N. Brunel. Firing rate of the noisy quadratic integrate-and-fire neuron. *Neural Computation*, 15:2281–2306, 2003.
- [139] P.E. Latham, B.J. Richmond, P.G. Nelson, and S. Nirenberg. Intrinsic dynamics in neuronal networks. I. Theory. *J. Neurophysiol.*, 83:808–827, 2000.
- [140] E.L. Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, 1959.
- [141] T.J. Lewis and J. Rinzel. Dynamics of spiking neurons connected by both inhibitory and electrical coupling. *J. Comp. Neurosci.*, 14:283–309, 2003.
- [142] J.E. Lisman, J.M. Fellous, and X.-J. Wang. A role for NMDA-receptor channels in working memory. *Nat. Neurosci.*, 1:273–275, 1998.
- [143] J.E. Lisman and M.A. Idiart. Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science*, 267:1512–1515, 1995.
- [144] Y.H. Liu and X.-J. Wang. Spike-frequency adaptation of a generalized integrate-and fire model neuron. *J. Comput. Neurosci.*, 10:25–45, 2001.
- [145] Y. Loewenstein, S. Mahon, P. Chadderton, K. Kitamura, H. Sompolinsky, Y. Yarom, and M. Hausser. Bistability of cerebellar purkinje cells modulated by sensory stimulation. *Nat. Neurosci.*, 8:202–211, 2005.
- [146] W.T. Maddox and C.J. Bohil. Base-rate and payoff effects in multidimensional perceptual categorization. *J. Experimental Psychology*, 24 (6):1459–1482, 1998.
- [147] I.G. Malkin. *Methods of Poincaré and Linstedt in the Theory of Nonlinear Oscillations*. Gostexisdat, Moscow, 1949. In Russian.
- [148] I.G. Malkin. *Some Problems in Nonlinear Oscillation Theory*. Gostexisdat, Moscow, 1956. In Russian.
- [149] E. Marder. Motor pattern generation. *Curr. Opin. Neurobiol.*, 10 (6):691–698, 2000.

- [150] E. Marder and D. Bucher. Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. *Annu. Rev. Physiol.*, 69:291–316, 2007.
- [151] A. Mason and A. Larkman. Correlations between morphology and electrophysiology of pyramidal neurons in slices of rat visual cortex. ii. electrophysiology. *J. Neurosci.*, 10:1415–1428, 1990.
- [152] M. Mattia and S. Fusi. Modeling networks with VLSI (linear) integrate-and-fire neurons. *Losanna, Switzerland: Proceedings of the 7th international conference on artificial neural networks*, 1997.
- [153] D.A. McCormick, B.W. Connors, J.W. Lighthall, and D.A. Prince. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *J. Neurophysiol.*, 54:782–806, 1985.
- [154] T. McMillen and P. Holmes. The dynamics of choice among multiple alternatives. *J. Math. Psych.*, 50:30–57, 2006.
- [155] T. McMillen and P. Holmes. An elastic rod model for anguilliform swimming. *J Math. Biol.*, 53:843–866, 2006.
- [156] T. McMillen, T.L. Williams, and P. Holmes. Nonlinear muscles, passive viscoelasticity and body taper conspire to create neuro-mechanical phase lags in anguilliform swimmers. *PLoS Comput. Biol.*, 4 (8):e1000157, 2008.
- [157] J. Myung and J.R. Busemeyer. Criterion learning in a deferred decision-making task. *American J. of Psychology*, 102 (1):1–16, 1989.
- [158] J.S. Nagumo, S. Arimoto, and S. Yoshizawa. An active pulse transmission line simulating a nerve axon. *Proc. IRE*, 50:2061–2070, 19.
- [159] T. Netoff, M.A. Schwemmer, and T.J. Lewis. Experimentally estimating phase response curves of neurons: theoretical and practical issues. In N.W. Schultheiss, A. Prinz, and R.J. Butera, editors, *PRCs in Neuroscience: Theory, Experiment and Analysis*, pages 95–129. Springer Verlag, 2012.
- [160] J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London Ser. A.*, 231:289–337, 1933.
- [161] H. Okamoto and T. Fukai. Physiologically realistic modelling of a mechanism for neural representation of intervals of time. *Biosystems*, 68:229–233, 2003.
- [162] H. Okamoto, Y. Isomura, M. Takada, and T. Fukai. Temporal integration by stochastic recurrent network dynamics with bimodal neurons. *J. Neurophysiol.*, 97:3859–3867, 2007.
- [163] K.G. Pearson. Central programming and reflex control of walking in the cockroach. *J. Exp. Biol.*, 56:173–193, 1972.
- [164] K.G. Pearson. Motor systems. *Curr. Opin. Neurobiol.*, 10:649–654, 2000.
- [165] K.G. Pearson and J.F. Iles. Nervous mechanisms underlying intersegmental co-ordination of leg movements during walking in the cockroach. *J. Exp. Biol.*, 58:725–744, 1973.

- [166] H.J. Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica*, 13:1–270, 1890.
- [167] H.J. Poincaré. *Les méthodes nouvelles de la mécanique celeste, Vols 1-3*. Gauthiers-Villars, Paris, 1892,1893,1899.
- [168] S.A. Prescott and Y. De Koninck. Gain control of firing rate by shunting inhibition: roles of synaptic noise and dendritic saturation. *Proc. Natl. Acad. Sci. USA*, 100:2076–2081, 2003.
- [169] A.A. Prinz, L.F. Abbott, and E. Marder. The dynamics clamp comes of age. *Trends Neurosci.*, 27:218–224, 2004.
- [170] J. Proctor and P. Holmes. Reflexes and preflexes: On the role of sensory feedback on rhythmic patterns in legged locomotion. *Biol. Cybern.*, 2:513–531, 2010.
- [171] J. Proctor, R.P. Kukillaya, and P. Holmes. A phase-reduced neuro-mechanical model for insect locomotion: feed-forward stability and proprioceptive feedback. *Phil. Trans. R. Soc. A*, 368:5087–5104, 2010.
- [172] R. Ratcliff. A theory of memory retrieval. *Psych. Rev.*, 85:59–108, 1978.
- [173] R. Ratcliff, A. Cherian, and M.A. Segraves. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two choice decisions. *J. Neurophysiol.*, 90:1392–1407, 2003.
- [174] R. Ratcliff, T. Van Zandt, and G. McKoon. Connectionist and diffusion models of reaction time. *Psych. Rev.*, 106 (2):261–300, 1999.
- [175] A. Rauch and S. Fusi. Neocortical pyramidal cells respond as integrate-and-fire neurons to in vivo-like input currents. *J. Neurophysiol.*, 90:1598–1612, 2003.
- [176] A. Renart, N. Brunel, and X.-J. Wang. Mean-field theory of recurrent cortical networks: from irregularly spiking neurons to working memory. In J. Feng, editor, *Computational Neuroscience: A comprehensive approach*. CRC Press, Boca Raton, 2003.
- [177] L. M. Ricciardi. *Diffusion Processes and Related Topics in Biology*. Springer Verlag, Berlin, 1977.
- [178] M.J.E. Richardson, N. Brunel, and V. Hakim. From subthreshold to firing-rate resonance. *J. Neurophysiol.*, 89:2538–2554, 2003.
- [179] F. Rieke, W. Bialek, D. Warland, and R. De Ruyter van Steveninck. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA, 1997.
- [180] J. Rinzel. Excitation dynamics: Insights from simplified membrane models. *Fed. Proc.*, 37:2944–2946, 1985.
- [181] A.E. Rorie, J. Gao, J.L. McClelland, and W.T. Newsome. Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. *PLoS One*, 5 (2):e9308, 2010.
- [182] 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.

- [183] S. Ross. *A First Course in Probability*. Prentice Hall, Upper Saddle River, New Jersey, 2002.
- [184] A. Roxin and A. Ledberg. Neurobiological models of two-choice decision making can be reduced to a one-dimensional nonlinear diffusion equation. *PLoS Comput. Biol.*, 4(3):e10000146, 2008.
- [185] D.E. Rumelhart and J.L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge, MA., 1986.
- [186] J.D. Schall. Neural basis of deciding, choosing and acting. *Nature Reviews in Neuroscience*, 2:33–42, 2001.
- [187] J. Schmitt, M. Garcia, C. Razo, P. Holmes, and R.J. Full. Dynamics and stability of legged locomotion in the horizontal plane: A test case using insects. *Biol. Cybern.*, 86(5):343–353, 2002.
- [188] J. Schmitt and P. Holmes. Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane – I. Theory. *Biol. Cybern.*, 83(6):501–515, 2000a.
- [189] J. Schmitt and P. Holmes. Mechanical models for insect locomotion: Dynamics and stability in the horizontal plane – II. Application. *Biol. Cybern.*, 83(6):517–527, 2000b.
- [190] J. Schmitt and P. Holmes. Mechanical models for insect locomotion: Active muscles and energy losses. *Biol. Cybern.*, 89(1):43–55, 2003.
- [191] N.W. Schultheiss, A. Prinz, and R.J. Butera, editors. *PRCs in Neuroscience: Theory, Experiment and Analysis*. Springer Verlag, New York, 2012. Springer Series in Computational Neuroscience, Vol. 6.
- [192] M.N. Shadlen and W.T. Newsome. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.*, 4:569–579, 1994.
- [193] M.N. Shadlen and W.T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.*, 18:3780–3896, 1998.
- [194] 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.
- [195] C.E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423; 623–656, 1948. Reprinted in Shannon and Weaver (1949) [196].
- [196] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL., 1949.
- [197] A. Sherman, J. Rinzel, and J. Keizer. Emergence of organized bursting in clusters of pancreatic β -cells by channel sharing. *Biophys. J.*, 54:411–425, 1988.
- [198] O. Shriki, D. Hansel, and H. Sompolinsky. Rate models for conductance-based cortical neuronal networks. *Neural Computation*, 15:1809–1841, 2003.
- [199] Y. Shu, A. Hasenstaub, and D.A. McCormick. Turning on and off recurrent balanced cortical activity. *Nature*, 423:288–293, 2003.

- [200] P. Simen, D. Contreras, C. Buck, P. Hu, P. Holmes, and J.D. Cohen. Reward rate optimization in two-alternative decision making: Empirical tests of theoretical predictions. *J. Exp. Psych. Human Perception and Performance*, 35 (6):1865–1897, 2009.
- [201] P.A. Simen, J.D. Cohen, and P. Holmes. Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, 19:1013–1026, 2006.
- [202] J.D. Skufka. Analysis still matters: A surprising instance of failure of Runge-Kutta-Felberg ODE solvers. *SIAM Review*, 46 (4):729–737, 2004.
- [203] P.L. Smith and R. Ratcliff. Psychology and neurobiology of simple decisions. *Trends in Neurosci.*, 27 (3):161–168, 2004.
- [204] R. Stein. A theoretical analysis of neuronal variability. *Biophys. J.*, 5:173–194, 1965.
- [205] E.A. Stern, A.E. Kincaid, and C.J. Wilson. Spontaneous subthreshold membrane potential fluctuations and action potential variability of rat corticostriatal and striatal neurons in vivo. *J. Neurophysiol.*, 77:1697–1715, 1997.
- [206] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, New York, 2002. Third edition.
- [207] A.M. Stuart and A.R. Humphries. *Dynamical Systems and Numerical Analysis*. Cambridge University Press, Cambridge, U.K., 1996.
- [208] G. Svirskis, R. Dodla, and J. Rinzel. Subthreshold outward currents enhance temporal integration in auditory neurons. *Biol. Cybern.*, 89:333–340, 2003.
- [209] A. Treves. Mean-field analysis of neuronal spike dynamics. *Network*, 4:259–284, 1993.
- [210] T.W. Troyer and K.D. Miller. Physiological gain leads to high isi variability in a simple model of a cortical regular spiking cell. *Neural Computation*, 9:971–983, 1997.
- [211] M. Tsodyks and T. Sejnowski. Rapid switching in balanced cortical network models. *Network*, 6:1–14, 1995.
- [212] E.D. Tytell, P. Holmes, and A.H. Cohen. Spikes alone do not behavior make: why neuroscience needs biomechanics. *Curr. Opin. Neurobiol.*, 21:816–822, 2011.
- [213] E.D. Tytell, C.-Y. Hsu, T.L. Williams, A.H. Cohen, and L.J. Fauci. Interactions between body stiffness, muscle activation, and fluid environment in a neuromechanical model of lamprey swimming. *Proc. Natl. Acad. Sci.*, 107(46):19832–19837, 2010.
- [214] M. Usher and J.L. McClelland. On the time course of perceptual choice: The leaky competing accumulator model. *Psych. Rev.*, 108:550–592, 2001.
- [215] C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724–1726, 1996.
- [216] P.L. Várkonyi, T. Kiemel, K. Hoffman, A.H. Cohen, and P. Holmes. On the derivation and tuning of phase oscillator models for lamprey central pattern generators. *J. Comput. Neurosci.*, 25:245–261, 2008.

- [217] J. von Neumann. *The Computer and the Brain*. Yale University Press, New Haven, CT, 1958. 2nd Edition, with a foreword by P.M. and P.S. Churchland, 2000.
- [218] A. Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.
- [219] A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, 19:326–339, 1948.
- [220] P. Wallen, O. Ekeberg, A. Lansner, L. Brodin, H. Traven, and S. Grillner. A computer-based model for realistic simulations of neural networks. II. The segmental network generating locomotor rhythmicity in the lamprey. *J. Neurophysiol.*, 68:1939–1950, 1992.
- [221] W.A. Wallis. The Statistical Research Group, 1942-1945. *J. Amer. Statist. Assoc.*, 75:320–330, 1980.
- [222] H. Wang, G.G. Stradtman 3rd, X.-J. Wang, and W.-J. Gao. A specialized NMDA receptor function in layer 5 recurrent microcircuitry of the adult rat prefrontal cortex. *Proc. Natl. Acad. Sci. USA*, 105:16791–16796, 2008.
- [223] X.-J. Wang. Fast burst firing and short-term synaptic plasticity: a model of neocortical chattering neurons. *Neuroscience*, 89:347–362, 1999.
- [224] X.-J. Wang. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.*, 19:9587–9603, 1999.
- [225] X.-J. Wang. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36:955–968, 2002.
- [226] X.-J. Wang. Decision making in recurrent neuronal circuits. *Neuron*, 60:215–234, 2008.
- [227] X.-J. Wang and G. Buzsaki. Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model. *J. Neurosci.*, 16:6402–6413, 1996.
- [228] R. Weinstock. *Calculus of Variations*. McGraw-Hill, New York, 1952. Reprinted by Dover Publications Inc., New York, 1974.
- [229] N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. M.I.T. Press, Cambridge, MA, 1948. 2nd Edition, 1961.
- [230] T.L. Williams. Phase coupling by synaptic spread in chains of coupled neuronal oscillators. *Science*, 258:662–665, 1992.
- [231] H. Wilson. *Spikes, Decisions and Actions: The Dynamical Foundations of Neuroscience*. Oxford University Press, Oxford, U.K., 1999.
- [232] H. Wilson and J. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12:1–24, 1972.
- [233] H. Wilson and J. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13:55–80, 1973.
- [234] A.T. Winfree. *The Geometry of Biological Time*. Springer Verlag, New York, 2001. Second Edition.

- [235] M. Winograd, A. Destexhe, and M.V. Sanchez-Vives. Hyperpolarization-activated graded persistent activity in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA*, 105:7298–7303, 2008.
- [236] K.F. Wong and X.-J. Wang. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.*, 26 (4):1314–1328, 2006.
- [237] M. Zacksenhouse, P. Holmes, and R. Bogacz. Robust versus optimal strategies for two-alternative forced choice tasks. *J. Math. Psych.*, 54:230–246, 2010.