A Search For Principles of Basal Ganglia Function

by

Bryan Tripp

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The basal ganglia are a group of subcortical nuclei that contain about 100 million neurons in humans. Different modes of basal ganglia dysfunction lead to Parkinson's disease and Huntington's disease, which have debilitating motor and cognitive symptoms. However, despite intensive study, both the internal computational mechanisms of the basal ganglia, and their contribution to normal brain function, have been elusive. The goal of this thesis is to identify basic principles that underlie basal ganglia function, with a focus on signal representation, computation, dynamics, and plasticity.

This process begins with a review of two current hypotheses of normal basal ganglia function, one being that they automatically select actions on the basis of past reinforcement, and the other that they compress cortical signals that tend to occur in conjunction with reinforcement. It is argued that a wide range of experimental data are consistent with these mechanisms operating in series, and that in this configuration, compression makes selection practical in natural environments. Although experimental work is outside the present scope, an experimental means of testing this proposal in the future is suggested.

The remainder of the thesis builds on Eliasmith & Anderson's Neural Engineering Framework (NEF), which provides an integrated theoretical account of computation, representation, and dynamics in large neural circuits. The NEF provides considerable insight into basal ganglia function, but its explanatory power is potentially limited by two assumptions that the basal ganglia violate. First, like most large-network models, the NEF assumes that neurons integrate multiple synaptic inputs in a linear manner. However, synaptic integration in the basal ganglia is nonlinear in several respects. Three modes of nonlinearity are examined, including nonlinear interactions between dendritic branches, nonlinear integration within terminal branches, and nonlinear conductance-current relationships. The first mode is shown to affect neuron tuning. The other two modes are shown to enable alternative computational mechanisms that facilitate learning, and make computation more flexible, respectively.

Secondly, while the NEF assumes that the feedforward dynamics of individual neurons are dominated by the dynamics of post-synaptic current, many basal ganglia neurons also exhibit prominent spike-generation dynamics, including adaptation, bursting, and hysterses. Of these, it is shown that the NEF theory of network dynamics applies fairly directly to certain cases of firing-rate adaptation. However, more complex dynamics, including nonlinear dynamics that are diverse across a population, can be described using the NEF equations for representation. In particular, a neuron's response can be characterized in terms of a more complex function that extends over both present and past inputs. It is therefore straightforward to apply NEF methods to interpret the effects of complex cell dynamics at the network level.

The role of spike timing in basal ganglia function is also examined. Although the basal ganglia have been interpreted in the past to perform computations on

the basis of mean firing rates (over windows of tens or hundreds of milliseconds) it has recently become clear that patterns of spikes on finer timescales are also functionally relevant. Past work has shown that precise spike times in sensory systems contain stimulus-related information, but there has been little study of how post-synaptic neurons might use this information. It is shown that essentially any neuron can use this information to perform flexible computations, and that these computations do not require spike timing that is very precise. As a consequence, irregular and highly-variable firing patterns can drive behaviour with which they have no detectable correlation.

Most of the projection neurons in the basal ganglia are inhibitory, and the effect of one nucleus on another is classically interpreted as subtractive or divisive. Theoretically, very flexible computations can be performed within a projection if each presynaptic neuron can both excite and inhibit its targets, but this is hardly ever the case physiologically. However, it is shown here that equivalent computational flexibility is supported by inhibitory projections in the basal ganglia, as a simple consequence of inhibitory collaterals in the target nuclei.

Finally, the relationship between population coding and synaptic plasticity is discussed. It is shown that Hebbian plasticity, in conjunction with lateral connections, determines both the dimension of the population code and the tuning of neuron responses within the coded space. These results permit a straightforward interpretation of the effects of synaptic plasticity on information processing at the network level.

Together with the NEF, these new results provide a rich set of theoretical principles through which the dominant physiological factors that affect basal ganglia function can be more clearly understood.

Acknowledgements

First, an NSERC scholarship was critical in allowing me to focus intensely on my work for long periods of time.

The PhD has been a pivotal experience for me. I was particularly fortunate in my choice of primary advisor, Chris Eliasmith. Chris is sharp, insightful, very supportive, and exciting to work with. I am also grateful to my co-advisor Dan Stashuk, who helped keep things on an even keel, and made several penetrating suggestions, despite having quite a different research focus. Thanks also to the other professors on my advisory committee, and in the Centre for Theoretical Neuroscience, particularly Brian Ingalls, Bill Hutchison, Paul Thagard, James Danckert, Sue-Ann Campbell, and very-particularly David Spafford, who always welcomed me into his lab. All of these people have been valuable role models, and from the beginning they treated me not so much like a student as a peer-to-be. Shu Wu was a tremendous help with the development of simulation software. Finally, it was a pleasure to share lab space with the many young geniuses that Chris attracted, including Terry Stewart, Ray Singh, Lloyd Elliot, James Martens, Chris Parisien, Jon Fishbein, Marc Hurwitz, Jean-Frederique Pasquale, John Conklin, Abninder Litt, Travis DeWolf, and many others.

Thanks also to my family. Lately we've sometimes taken on too much at once, and the parents and siblings have been there consistently to help us through. My dad in particular spent a lot of his time helping us keep the house in good shape. We re-wired the whole place last year, and if it wasn't for him I'd still be there, covered in plaster, with no thesis. Thanks to my artistic sister Carolyn for Figure 1.1. Most importantly, my wife Andrea has been a wonderful, understanding, and cherished companion. These first few years of our marriage were great, and I am sorry to see them go. We've come a long way together.

Contents

Chapter 1

Introduction

This introductory chapter briefly describes the basal ganglia, and introduces the Albin/DeLong model, a conceptual model of basal ganglia pathology. This model has been a key tool for understanding basal ganglia function for two decades, and modern basal ganglia research can be understood in terms its limitations. Building on this research, and on developments in other areas of theoretical neuroscience, the goal of this thesis is to articulate basic principles that govern information processing in these nuclei.

1.1 Basal Ganglia

The basal ganglia are a densely interconnected group of subcortical nuclei, including the striatum, the substantia nigra, the globus pallidus, and the subthalamic nucleus. The basal ganglia (BG) lie beneath the cerebral cortex and around the thalamus, and are strongly connected with both structures.

Major intrinsic and extrinsic connections of the basal ganglia are shown in Figure 1.1. Much of this connectivity can be understood in terms of feedback loops with the cortex. The cortex projects massively to the striatum, which is the main input structure of the basal ganglia. The main output nuclei are the internal segment of the globus pallidus (GPi), and the substantia nigra pars reticulata (SNr). These project to the thalamus, which projects to the cortex, completing the loop. There are multiple paths from the striatum to the output nuclei.

An unusual feature of the basal ganglia is that the projection neurons of most nuclei contain the neurotransmitter GABA, and have an inhibitory effect on their targets. The shortest path through the basal ganglia (called the "direct pathway" or "primary axis") consists of the projection from the striatum to the GPi, and from the GPi to the thalamus. Both of these projections are inhibitory, and the GPi neurons are tonically active. Thus cortical excitation of the striatum inhibits GPi neurons, which then fire more slowly, and inhibit thalamic neurons less than

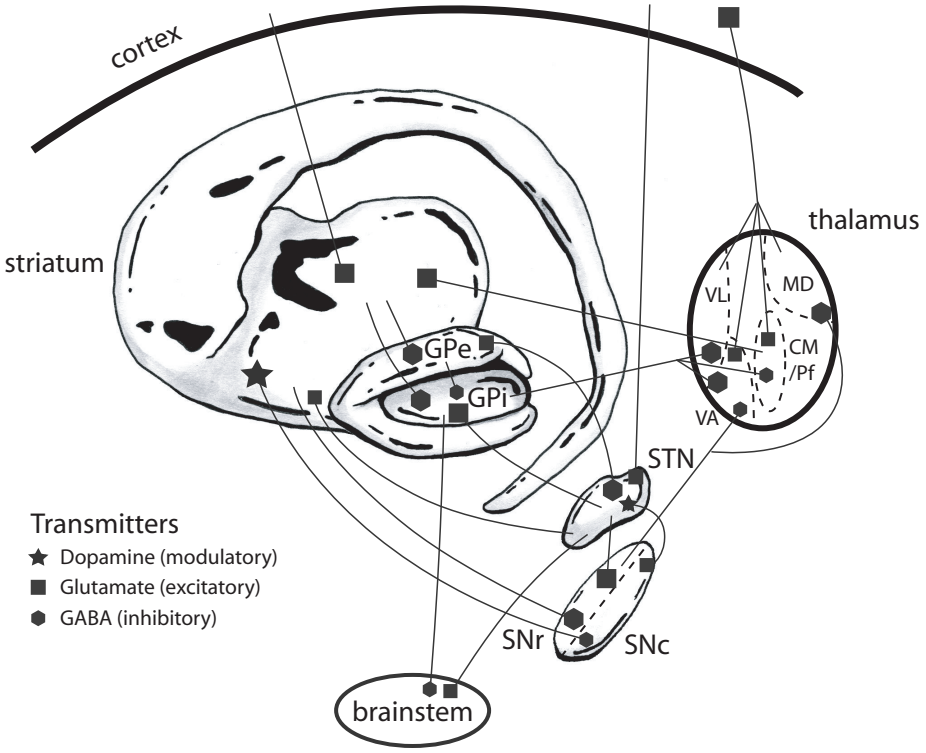

Figure 1.1: Major connections of the cortico-basal ganglia-thalamo-cortical system (abstracted from various sources including [293, 294, 173, 387]). Abbreviations: CM: centromedian nucleus of the thalamus; Pf: parafascicular nucleus; VA: ventral anterior nucleus; VL: ventral lateral nucleus; MD: medio-dorsal nucleus; SNc: substantia nigra pars compacta; GPe: external globus pallidus; GPi: internal globus pallidus; SNr: substantia nigra pars reticulata; STN: subthalamic nucleus. Arrowhead shapes indicate the main neurotransmitter of each projection (see key bottom left). Connections that are particularly massive are shown with larger arrowheads. Numerous minor connections are omitted for simplicity. The substantia nigra, globus pallidus, and STN are almost completely composed of neurons that project to other nuclei. However, the striatum contains several types of locally-projecting interneurons (see [387] for review). All of the thalamic nuclei shown here contain GABAergic local circuit neurons.

usual. Thalamic neurons can then fire more quickly, and excite the cortex more. Direct-pathway striatal activity is thus said to disinhibit the thalamus.

In addition to the direct pathway there are a number of side loops. One of these is the "indirect pathway", which consists of the inhibitory projection from the striatum to the external segment of the globus pallidus (GPe), the inhibitory projection from GPe to the subthalamic nucleus (STN), and the excitatory projection from STN to the output nuclei. Striatal activity has a net inhibitory effect on the thalamus along this pathway. The direct and indirect pathways arise from distinct groups of projection neurons in the striatum. Both groups are morphologically similar, and contain GABA. However the two groups are distinct in that those of the direct pathway contain dopamine receptors of the D1 family and cotransmit substance P and dynorphin, while those of the indirect pathway contain D2-family receptors and cotransmit enkephalin. The D1 and D2 receptors mediate different post-synaptic effects.

As shown in Figure 1.1, there are other major paths through the basal ganglia. Of particular interest is the inhibitory projection from GPe to GPi. Axons of this projection form dense bundles of inhibitory synaptic contacts around the cell bodies of the target GPi neurons, suggesting a strong coupling [294]. Another pathway that has received much attention recently is the direct projection from the cortex to the STN. Cortical influences reach the basal ganglia output structures more quickly via this path than via the more massive path through the striatum [272].

1.1.1 Basal Ganglia Function

It is not clear what the basal ganglia do in the healthy brain. A variety of suggestions have been made, including roles in reinforcement learning and sequence production. Mink [262] and others have suggested that when the cortex selects and executes an action, the basal ganglia act to inhibit competing actions and facilitate the selected action. Since the loops that connect with motor and non-motor cortical areas are similar in many respects, it has been proposed that influence of the basal ganglia on motor and non-motor functions may be analogous. The basal ganglia may also have similar influences on different motor processes, so there is particular interest in studying their role in motor systems that are relatively well characterized, such as the oculomotor system [156, 115] and the circuits controlling vocalization in songbirds [105]. Evidence related to normal basal ganglia function is reviewed at length in Chapter 2.

1.1.2 Basal Ganglia Dysfunction

Unfortunately, it is much more clear what the basal ganglia do when they are not working properly.

Parkinson's Disease

Much of the practical motivation for studying the basal ganglia is due to Parkinson's Disease (PD), which causes focused degeneration of the dopamine-releasing neurons of the substantia nigra compacta (SNc; see Figure 1.1). PD is the second-most common neuro-degenerative condition (after Alzheimer's Disease). The condition slowly progresses in severity and resistance to treatment. Patients develop striking motor deficits, including difficulty initiating movement (akinesia), slowness of movement (bradykinesia), resting tremor, rigidity, shuffling gait, and balance impairments. Particularly impacted are sequential or simultaneous coordinated movements [333, 47, 48], and movements demanding high precision [10]. Many patients medicated with levodopa develop involuntary movements after some period of treatment [139, 112]. Patients also suffer from cognitive impairments, and treatments have cognitive side-effects.

PD has diverse causes, including inherited mutations and exposure to environmental toxins. However, all cases of PD appear to involve dysfunction of α-synuclein, a protein which is thought to be involved in neurotransmitter release. High levels of α-synuclein are expressed throughout the brain [339], so it is not immediately obvious why damage in PD should be concentrated in the SNc. However, some byproducts of dopamine metabolism are highly reactive, and it is possible that α-synuclein either enhances their production or interferes with their inactivation [307].

Much of the scientific study of Parkinsonism has been based on animal models. Primates can be rendered Parkinsonian by administration of the toxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP), which selectively affects dopaminergic neurons. Other animal models include rats lesioned with 6-hydroxy-dopamine, and various genetic manipulations, particularly in mice. Molecular mechanisms can also be studied in insects (e.g. [116]).

Treatment of PD typically begins with drugs that compensate for the loss of dopamine-releasing neurons [268]. Administration of levodopa, a chemical precursor of dopamine, enhances the dopamine output of surviving cells. Other drugs act directly on dopamine receptors. Drug-based treatments become less effective as the disease progresses, and must often be complemented by surgical interventions. Surgery can involve lesioning parts of the basal ganglia which are believed in later disease stages to be generating dysfunctional output. Chronic high-frequency electrical stimulation of the same areas is an effective alternative, with advantages over lesioning, in that the stimulation can be fine-tuned after the operation, or discontinued if necessary. GPi stimulation improves drug tolerance by reducing drug-induced dyskinesias; STN stimulation improves other major PD symptoms, allowing lower drug dosages [211, 269]. There has also been promising research related to other potential interventions, such as cell therapy [228] and immunization [244].

Other Basal Ganglia Disorders

Huntington's disease involves progressive loss of striatal neurons. In contrast with PD, the primary sign in early stages of Huntington's is choreic movement (involuntary, writhing, vaguely dance-like movement). Bradykinesia develops with disease progression, although it is different from the bradykinesia of PD [50]. Non-motor symptoms (e.g. personality changes) are also different from those of PD.

Dystonia is a family of conditions that cause involuntary movement and abnormal postures. Some cases arise from an identifiable injury, which is often to the basal ganglia [217, 71], and some have no obvious pathology. Some authors have reported changes in firing patterns and reduction in mean firing rates in GPi, although this may be an artefact of general surgical anaesthesia [171]. Lesion [375] and electrical stimulation [374, 171] of the GPi have been used to treat dystonia.

Tourette syndrome is also thought to be related to basal ganglia dysfunction, although the exact cause is unclear [263]. Symptoms include motor tics, and uncontrolled stereotyped behaviours including (in a minority of cases) verbal outbursts. There is also evidence of basal ganglia involvement in schizophrenia, obsessive-compulsive disorder, and attention-deficit-hyperactivity disorder.

1.2 Albin/DeLong Model

An elegant conceptual model of basal ganglia motor pathology (Figure 1.2) was introduced in influential reviews by Albin, Young & Penney [11] and DeLong [96]. This model is consistent with key points of basal ganglia anatomy and the major neurotransmitter effects, and accounts (qualitatively) for many symptoms of Parkinson's disease, Huntington's disease, and ballism (involuntary ballistic movement). The model is also consistent with new observations that emerged around the same time, such as the overactivity of STN in PD, relief of PD symptoms with STN lesion, and the different effects of D1 and D2 receptor activation.

The model emphasizes the opposing effects of the direct and indirect pathways. The direct pathway is characterized as promoting cortical activity, including particularly movement-related activity in the supplementary motor area of the cortex. In this pathway, striatal neurons disinhibit thalamic neurons (via the GPi and SNr), which in turn excite cortical cells. In contrast, the indirect pathway is proposed to inhibit cortical activity. The striatal neurons at the head of this pathway disinhibit STN via GPe, causing the excitatory STN neurons to increase activity in GPi and SNr, which inhibits the thalamus. It is not specified how the two pathways might interact to influence movement. Various possibilities are consistent with the model, including that the direct pathway encourages a desired movement while the indirect pathway discourages others, and that balance between the direct and indirect pathways scales movements in some way [383].

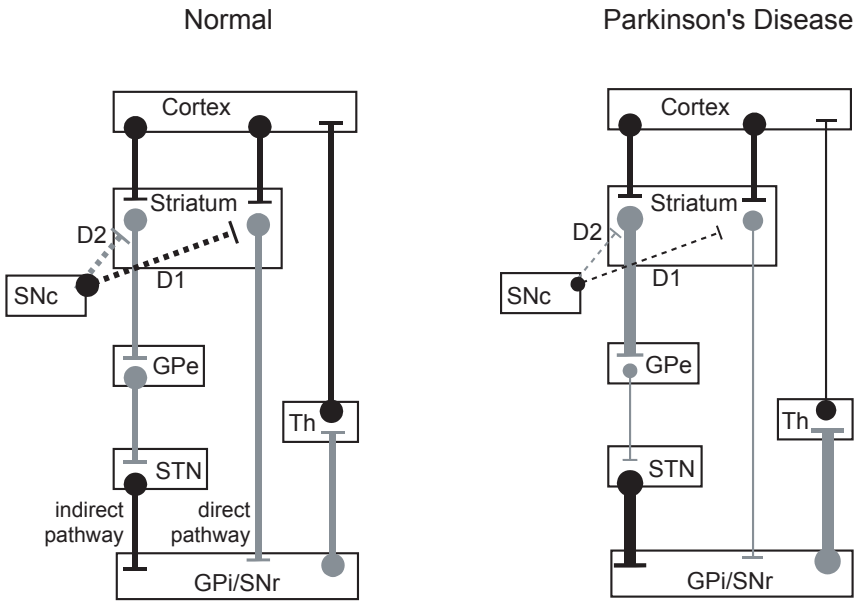

Figure 1.2: The Albin/DeLong model's account of activity in the major basal ganglia projections, in health and Parkinson's disease. Th=thalamus; other abbreviations as in text. Dark lines indicate excitatory connections, light lines indicate inhibitory connections, and dashed lines represent dopamine-containing connections. Flat ends are pre-synaptic and round ends are post-synaptic. On the right, relative changes in activity in PD are indicated by line thickness (i.e. thinner line indicates less activity than normal; thicker indicates more activity). Adapted from [11] and [96].

Recall that the striatal neurons at the heads of the direct and indirect pathways contain different dopamine receptors. In the Albin/DeLong model, dopamine increases activity in the striatal neurons of the direct pathway (via D1 receptors) and decreases activity of the striatal neurons of the indirect pathway (via D2 receptors). Thus with depleted dopamine in PD, the balance shifts in favour of the indirect pathway, consistent with symptoms of slowed movement and difficulty initiating movement.

D2-containing neurons are preferentially destroyed in the early stages of Huntington's disease. The model can therefore account for associated choreic movements, on the basis of reduced activity in the indirect pathway. The differential effect of dopamine via D1 and D2 receptors is also consistent with the amelioration of Huntington's disease symptoms by D2 receptor antagonists. Ballism resembles an acute version of chorea. Ballism is normally caused by destruction of the STN, so it is also accounted for by the model, as a decrease in inhibitory BG output arising from a disruption of the indirect pathway.

1.2.1 Limitations of the Albin/DeLong Model

Several limitations arise from the model's simplicity. It omits several major anatomical connections, including the direct projection from GPe to GPi, the cortico-subthalamic projection, and the thalamostriatal projection. It also overstates the distinctions between direct and indirect pathways. Many striatal axons branch to both GPe and GPi [291, 231], and D1 and D2 receptors are co-localized (although in different concentrations) on most striatal projection neurons (medium-spiny neurons) [8].

Furthermore, since the model is qualitative, only vague predictions are made as to the activities of various neuronal populations. It ignores intrinsic properties of the nuclei, such as neuronal diversity, connections intrinsic to each nucleus, and membrane dynamics. It is also simplistic in its representation of extrinsic connections. For example the only predictions made regarding the BG influence on the cortex are that it will be greater or lesser in various circumstances. So while the model is generally consistent with many disease symptoms, it does not predict them with precision or in detail. It does not account for the Parkinsonian tremor symptom.

Moreover, predictions of the model conflict with several recent observations. Although there were early indications that GPe activity was reduced in PD, more recent studies have not provided strong support. GPe firing rates in dopamine-lesioned rats and monkeys are reduced only moderately (and may return to normal in a matter of weeks), and long-term metabolic activity in GPe is generally unchanged [224]. Also, STN activity is not strongly correlated with GPe activity in the healthy brain [371], and the confirmed increase in STN activity in PD appears to have less to do with GPe than with other factors [224, 291]. Predicted differences in GPi firing rates between PD and Huntington's disease patients are not consistently observed [356, 347]. The model does accurately predict that STN or GPi lesion in PD would improve hypokinetic symptoms. However, in conflict with the model, hyperkinetic symptoms are also improved by these lesions. Also, thalamic lesions would be expected to worsen PD symptoms, but in fact improve them [283]. The model also does not account for the fact that high-frequency stimulation of the thalamus, pallidum, and STN produce similar clinical outcomes to lesions of these nuclei [46].

Even the existence of the indirect pathway through the STN is not beyond doubt. Some studies suggest that the parts of the STN that receive input from GPe project back primarily to GPe [294], while other results suggest that most STN neurons that project to GPe collateralize to GPi, and that these are not physically segregated from those that project exclusively to the output nuclei [326].

Despite these limitations, the fact that the Albin/DeLong model elegantly accounts for a number of major characteristics of basal ganglia activity and disease makes it an important reference point against which to compare more sophisticated models. It is also important to acknowledge the great impact the model has had, not

only in organizing research, but also in leading to lesion and stimulation treatments for advanced PD, which are the most effective options currently available.

1.3 Principles, Models, and Principled Models

The Albin/DeLong model is important background for any dicussion of the basal ganglia, but it is hardly the state of the art. Much more specific proposals have been made about the influence of the basal ganglia on the cortex, and sophisticated computational models have accompanied them. The next chapter will take a close look at some of these proposals, and further review will accompany each later chapter as needed.

In the mean time, the remainder of this chapter introduces the intended role of the present thesis in ongoing basal ganglia research. Broadly speaking, the goal of this thesis is to use engineering methods (e.g. signals and systems theory; numerical simulations) to achieve a better understanding of how the basal ganglia work. The obvious way to do this would be to develop a new computational model. However, while models play an important supporting role in this thesis, the main focus is on identifying more basic theoretical points. This section explains the rationale behind this emphasis.

1.3.1 Principles

The 20th century witnessed the discovery of many key principles of neuroscience, without which it would now be hard to imagine thinking about brain function at all. Some of the most fundamental are as follows:

1. Action potentials are all-or-nothing events; the information conveyed by an action potential consists of the time at which it occurs.
2. Rapid information flow at chemical synapses is uni-directional, from the presynaptic neuron to the post-synaptic neuron.
3. All of the chemical synapses originating from a single neuron release the same primary neuro-transmitter.

All of these principles are approximations, and there are exceptions. However, the approximations are close, and the exceptions are relatively rare. The brain can be understood much more clearly if it is considered in terms of principles like these, rather than as a massive list of individual cases. A principle distills a pattern of facts into more manageable form. If we want to understand the brain more clearly, we will have to unearth more principles.

1.3.2 Models

Models provide an additional, complementary aid to understanding. A conceptual model, like the Albin/DeLong model, organizes diverse observations into a simple framework that (very roughly) reflects the behaviour of a complex system.

Recently, *computational* models have provided additional rigour. The key factor that distinguishes computational models from conceptual ones is that computational models are expected to function. This means that no parameter can be left unspecified. Model development can therefore expose missing data, and/or lead to new ideas about which data are important, ultimately suggesting new experiments. It has even been argued that computational models are the ultimate goal of neuroscience [255]. This is because the brain is so complex that no matter how many facts about it are obtained through experiments, that there is no hope of accurately understanding these facts unless they are integrated in a rigorous and quantitative manner.

Computational models can be highly abstract. Abstract computational models are essentially well-specified conceptual models, and they provide a means of formalizing simple ideas about how a system works. At the other end of the spectrum, there is increasing interest in more sophisticated models that embrace complex physiological details. One reason for this is the realization that neurons are complex, noisy physical devices, which 1) may actually be incapable of behaving like elegant equations, and on the other hand, 2) can clearly do things that have yet to be described with equations. Another reason for this interest is that, ideally, realistic models can be experimented on, i.e. manipulated and measured much like real brains, but with far greater flexibly.

However, as these models become more sophisticated, they must also rely increasingly on tractable principles, or they run the risk of becoming incomprehensible themselves. Furthermore, only a tiny fraction of the brain's parameters can be measured (e.g. the human brain has about 100 trillion synapses, of varying strength), so sophisticated computational models are severely underconstrained. Less-principled models run a greater risk of fitting the available data without reflecting the essential structure of the corresponding system.

1.3.3 Principled Models

These considerations motivate Eliasmith & Anderson's [111] Neural Engineering Framework (NEF), from which this thesis draws heavily. At the core of the framework are three key theoretical principles (their wording):

1. Neural representations are defined by the combination of nonlinear encoding and weighted linear decoding.

2. Transformations of neural representations are functions of variables that are represented by neural populations. Transformations are determined using an alternately weighted linear decoding.

3. Neural dynamics are characterized by considering neural representations as control-theoretic state variables.

As discussed in detail in Chapter 3, the above principles and the associated equations provide a systematic way to develop models that integrate diverse experimental data, including high-level behavioural data and lower-level electrophysiological data.

So perhaps a good way to understand the basal ganglia would be to model them using the NEF. But there are two limitations inherent in this approach. The first is that despite the systematicity introduced by the framework, there remain many free parameters. It would be hard to determine whether a principled NEF model that fit a great deal of experimental data was unique in doing so.

However, a more fundamental limitation is that it is not clear that the NEF principles encompass all of the features of basal ganglia circuits that are relevant to their behaviour. For example, a closer look at NEF's treatment of network dynamics (the third principle, above) reveals assumptions which, strictly speaking, do not hold for many basal ganglia neurons.

Such points of divergence between the NEF and the basal ganglia (described in more detail in Chapter 3) have become a major focus of this thesis. Where the above principals either diverge from reality or do not clearly relate to it, one could reasonably choose to either ignore the differences, or to build a less principled model that respects them. However, it is probably more useful in the long term to set the task of modelling aside, and to search in these mismatches for additional principles.

1.4 Principles of Basal Ganglia Function

The principles that have emerged from this work are listed below, in the order in which they appear in the following chapters. As principles go, these are relatively humble and special-purpose. However, they all contribute to a richer understanding of the computational capacities of the basal ganglia networks.

1. *Dimensionality reduction is an effective precursor for action selection, and may precede action selection in feedforward basal ganglia circuits.* (Section 2.4) Dimensionality reduction and action selection are currently alternative hypotheses regarding basal ganglia function, each of which agrees more closely with a different subset of the experimental data. This section illustrates that the hypotheses are compatible, and discusses potential functional advantages of their combination in a single system.

2. *Nonlinear dendrites can perform computations in the absence of presynaptic tuning-curve diversity.* (Section 4.4) This is a more general point about computation in large networks. Network models often assume that neurons combine synaptic inputs in a near-linear manner, but there is evidence to the contrary. With linear synaptic integation, computation in a network is constrained by the range of diversity in the responses of different neurons. This section illustrates that dendritic nonlinearities can work around this requirement.

3. *Neurons can flexibly extract information from inputs that have neither time-varying spike rates nor precise spike timing.* (Section 5.3) Most models of the basal ganglia assume that information is carried in the rate at which neurons emit spikes, but there is growing evidence that the timing of individual spikes is important. This section illustrates that computation can proceed independently of spike rates, even when spike timing is highly variable.

4. *Inhibitory projections can calculate nonlinear, non-monotonic functions.* (Section 6.3) In contrast with the cortex, most projection neurons in the basal ganglia are inhibitory. Inhibition is usually equated with subtraction or division. This section shows that inhibition in a population code can subserve complex and flexible computations.

5. *Cell-intrinsic dynamics enable computations based on input history.* (Section 7.4) Most basal ganglia neurons exhibit nonlinear intrinsic firing dynamics, complicating quantitative theories of information representation. This section shows that representation in these neurons can be re-cast in terms of higher-dimensional representation of past and present inputs.

6. *Hebbian plasticity can set the dimension and tuning of a population code.* (Section 8.4) Neurons in the basal ganglia represent information in large groups, or populations. The details of this representation are established by synaptic plasticity, but the relationships between synaptic plasticity and population coding are not well understood. This section shows how local mechanisms of plasticity can determine two of the three key variables of a population code.

The above observations are theoretical principles, in the sense that each one of them could serve as a building block for a variety of basal ganglia models, either alone or in combination with others. The following chapters attempt to encourage this usage, by addressing each principle separately, rather than potentially obscuring them within a single complex model. However, they are each presented in the context of the more basic principles of the NEF, which provides a coherent underlying structure that clarifies the relationship between each of the above principles and the others.

Finally, a number of them also apply to networks outside the basal ganglia, particularly to cortical networks. The discussion in the following chapters is therefore

generalized to other systems wherever possible. This makes the focus broader at times than the basal ganglia, but this does not interfere with the goal of elucidating basal ganglia function. The commonalities between the basal ganglia and other networks are at least as informative as the differences.

Chapter 2

Action Selection vs. Dimensionality Reduction

Because the basal ganglia are central structures, far removed from sensory organs and muscles, identifying their function has not been straightforward. In the last decade, the dominant hypothesis has been that they automatically select context-appropriate actions. This hypothesis is broadly consistent with electrophysiological and anatomical evidence, behavioural data from reinforcement-learning experiments, and disease symptoms. An alternative hypothesis [37] is that the basal ganglia serve mainly to reduce the dimensionality of contextual information. This review chapter argues that the dimensionality reduction hypothesis is at least as consistent with anatomical and electrophysiological evidence as the action selection hypothesis, but that it fails to account for behavioural data. However, it is further argued that dimensionality reduction in the input stage of the basal ganglia would be an effective substrate for learning of action selection via reinforcement. From this perspective, the key outstanding question is: Where in the basal ganglia-cortical loop is the mapping from context to action performed? An experimental approach for addressing this question is proposed.

2.1 Introduction

The severity of symptoms in advanced Parkinson's disease makes it clear that the basal ganglia can have a major influence on cortical function. The Albin/DeLong model describes conditions in which this influence can become pathologically imbalanced. However, the nature of this influence is not completely clear. This is partly because the basal ganglia are central structures, several steps removed from both sensory organs and muscles.

The issue is further clouded by the fact that the functions of many of the cortical areas with which the basal ganglia are connected are themselves poorly-characterized or controversial. The relative exception is the motor areas of the

cortex. These are the best-understood of the basal ganglia targets, so one would expect the motor functions of the basal ganglia to be the most readily elucidated, and these functions have indeed received the most attention. However, even the precise function of the primary motor cortex (in which many projection neurons synapse directly onto primary motor neurons [109]) remains a subject of debate [362]. Nevertheless, its activity is closely correlated with movement execution (e.g. [129, 73]), suggesting a relatively direct role in driving movement.

In contrast to relatively direct role of the motor cortex in controlling the muscles, both the basal ganglia and cerebellum have relatively few descending projections, instead forming feedback loops with different cortical areas [167], including all of the motor areas [109]. Damage to the cerebellum can result in delayed movement initiation and poor coordination of different movement components. Different lesions and disorders of the basal ganglia have varied effects, but some of these effects relate clearly to high-level decisions between different courses of action [34]. Accordingly, it has been argued that basal ganglia anatomy and physiology are consistent with the role of automatically selecting actions that are adapted to the state of the animal's environment, and simultaneously suppressing the selection of alternative actions (e.g. [309, 52]). This mapping from context to action is thought to be established by reinformement learning (e.g. [165, 332]).

However, the basal ganglia do more than select motor actions. They also project to prefrontal areas of the cortex [16], suggesting that they play a role in cognitive activity (e.g. working memory, declarative memory retrieval, representation of goals). Accordingly, Parkinson's disease and Huntington's disease have cognitive symptoms as well as motor symptoms [108, 216], and the basal ganglia are implicated in several neuro-psychiatric conditions, including obsessive-compulsive disorder, attention-deficit-hyperactivity disorder, Tourette syndrome, and schizophrenia. Loops through the basal ganglia that affect cognitive and limbic areas of cortex have anatomy similar to the motor loops [16, 258], suggesting that the role of the basal ganglia in cognition may be analogous to that in motor control, i.e. that they select internal, cognitive actions, based on prior reinforcement [45, 309, 151].

Interestingly, this general form of the action-selection hypothesis has enjoyed great traction in the cognitive science community. A generalized selection mechanism would resemble a production system, i.e. a set of context-driven rules that guide behavior (a concept from artificial intelligence). One particularly successful cognitive modelling architecture, ACT-R [23], treats the basal ganglia as the anatomical substrate of a production system that guides communication between different cortical areas. ACT-R models are much farther removed from the underlying neurophysiology than models in neuroscience. But they model much more sophisticated behavior (e.g. eye movements in a complex air traffic control task [25]), with striking success. In these models, the basal ganglia select internal actions, such as retrieval of an item from declarative memory, or setting a subgoal in a complex task. Recent fMRI experiments generally support the timing of basal ganglia involvement predicted by these models [23] (but see [24]).

To summarize, the action-selection hypothesis has substantial experimental support, is plausibly related to the influence of basal ganglia on motor and non-motor frontal areas, and (although the neurophysiological details remain to be worked out) continues to make sense in models of complex behavior.

But challenges to this hypothesis remain. One puzzling issue is that fairly large lesions of the basal ganglia output nuclei as treatment for Parkinson's disease result in relatively mild impairment [283] (although lesioned patients do not move as smoothly or as quickly as healthy controls [200]). This suggests that if the basal ganglia are a selector of actions (or more generally, a production system), they are not the only one. Furthermore, lesions of the basal ganglia output nuclei in healthy experimental animals lead to noticeable deficits, but most of these are not obviously related to selection [164, 266, 100] (but see [223]).

More critically, the action selection hypothesis is not the only one that fits much of the data. In particular, the anatomy and electrophysiology of the basal ganglia are also largely consistent with the hypothesis that their main function is to reduce the dimensionality of contextual information, in a manner that emphasizes reward-relevant context [37].

The remainder of this chapter reviews the experimental evidence regarding basal ganglia function, and argues that the dimensionality-reduction hypothesis fits the anatomical and electrophysiological data at least as well as the action selection hypothesis, but that some of the behavioral data are more consistent with action selection. It is further argued that dimensionality reduction would provide an effective input stage for an action selection system, so that these hypotheses do not necessarily conflict. From this perspective, the key question is not whether the basal ganglia perform reduction or selection, but where in the basal ganglia pathways the map from context to action occurs.

2.2 Competing Theories

2.2.1 Action Selection

The action selection hypothesis is a natural elaboration of the Albin/DeLong model, in that one way to interpret the opposing effects of the direct and indirect pathways is that they respectively select some actions and supress others [99].

In an influential review, Mink [262] argued that 1) a central action-selection mechanism is necessary to prevent conflicting use of resources (e.g. muscles), 2) the basal ganglia are appropriately positioned for this function, since they receive convergent input from much of the cortex and project to the cortical areas that influence behaviour, 3) the selection function is consistent with internal circuitry of the basal ganglia, as well as basal ganglia disease symptoms (see also [264]), and electrophysiological and lesion experiments. Mink proposed that the hyper-direct pathway inhibits most candidate actions, through the relatively-divergent

projections of the STN [292], while stronger and more focused activity in the direct pathway disinhibits a single selected action. Redgrave and colleagues [309, 143] have made similar arguments, and developed a series of computational models based on this theme (e.g. [169]).

Multiple Selection Mechanisms

In order to evaluate the action-selection hypothesis, it will be important to distinguish between a few different modes of selection. Animals select actions in at least two ways. Some actions are goal-oriented, and sensitive to changes in both action-outcome contingencies and the reward value of outcomes. Others are stimulus driven, or habitual, and less sensitive to changing relationships between actions and rewards. There is recent evidence that the basal ganglia participate in both of these kinds of action selection. In particular, lesions of basal ganglia circuits that project to motor cortical areas reduce the propensity for stimulus-driven actions [395], while lesions of parallel basal ganglia circuits that project to prefrontal areas reduce the propensity for goal-oriented actions [397]. In the goal-oriented case, the basal ganglia probably do not select motor plans directly. They may instead contribute to the normal operation of prefrontal cognitive and working memory circuits [124], which among other things can make sophisticated action-related decisions, and subsequently influence the motor cortex *via* direct projections [363]. As mentioned above, the similarity between motor and non-motor basal ganglia circuits suggests that their contribution to prefrontal computations in this instance may relate to selection of cognitive steps toward the goal-oriented determination of motor plans.

Daw et al. [91] showed that stimulus-driven and goal-oriented selection mechanisms are optimal in different situations. They further proposed that these mechanisms normally operate in parallel, and that behaviour is dominated by the system that identifies a preferred action with greater confidence.[1] This group has also suggested that animals may sometimes simply duplicate their previous actions on the basis of declarative memory of similar situations [220]. This would constitute a third type of action selection. For humans, instruction-following arguably constitutes a distinct fourth mechanism, which is particularly important in situations that are novel to the individual [23]. In summary, there are multiple ways in which animals can select actions, and the basal ganglia may participate directly or indirectly in all of them. However, this review is concerned with the hypothesis that the basal ganglia perform selection directly, i.e. that their output biases selection of discrete alternative activity patterns in their cortical targets. Of the various mechanisms for selection of overt behaviours, stimulus-driven selection of motor actions is the one that relates most directly to this hypothesis.

[1] This proposal was motivated by the behaviour of lab rats in experimental settings, in which the rats have nothing better to do than optimize rewards. In more complex environments, practical advantages of the stimulus-driven mechanism (e.g. shorter reaction time; less interference with concurrent tasks) may take on more importance.

Reinforcement Learning

The models of Mink [262], Gurney et al. [143], and others focus on the mechanisms by which an animal selects from a list of discrete candidate actions over short time scales. But how does an animal come to prefer one action over others in the first place? It has long been recognized that animals' decisions are influenced by reinforcement [360], i.e. positive and negative outcomes associated with similar decisions in the animal's past experience. There are striking parallels between basal ganglia physiology and the mechanics of temporal-difference (TD) learning, a key reinforcement-learning algorithm from the machine-learning field [353].

TD learning is a means by which an agent can maximize the rewards it obtains in a complex environment. In contrast with other reinforcement learning algorithms, 1) it does not require complete *a priori* knowledge of the environment in order to learn an optimal context-dependent decision policy, and 2) it learns continuously, incorporating new information as it is encountered, even at times when a reward is neither expected nor available. One way to implement the TD algorithm is with a component that makes decisions (an "actor"), and another component that keeps track of the rewards associated with various environmental states, and uses this information to critique the actor's decisions (the "critic").

The output of the critic closely resembles the activity of dopamine neurons. For example, when monkeys self-initiate movement to a food reward, dopamine neurons are moderately active before movement, and fire a burst when the food is touched. Similar activity is observed when monkeys begin learning to associate a visual or auditory cue with an opportunity to obtain food by performing some task. However, when the task is well learned, dopamine neurons burst when the cue is presented, but not when the reward is obtained [316]. In fact there is a pause in tonic firing if the reward is unexpectedly withheld [158]. The firing of dopamine neurons in these experiments seems to reflect errors in the monkeys' expectations of reward, which is precisely the output of the critic in TD learning. The analogy between the actor-critic architecture and the basal ganglia is strengthened by the fact that the dopamine signal modulates cortico-striatal synaptic plasticity [312]. If the striatum represents context-specific actions, this modulation corresponds to the way in which the actor modifies its decision policy based on input from the critic. Several computational models have mapped the actor/critic TD architecture onto the basal ganglia in a more detailed manner [185].

2.2.2 Reinforcement-Driven Dimensionality Reduction

Although the action-selection hypothesis is consistent with a variety of experimental evidence (discussed further below), it also ignores some of the more striking anatomical and electrophysiological characteristics of the basal ganglia. One such characteristic is the funnel-like structure of the direct pathway – there are about one tenth as many striatal neurons as there are cortical neurons that project to the

striatum [403], and (depending on the species) a further reduction of 2-3 orders of magnitude between the striatum and the output nuclei [37, 387]. Another is that in contrast with the cortex [372], the firing activity of neighbouring basal ganglia neurons tends to be uncorrelated, or very weakly correlated (i.e. to have a fairly flat cross-correlation function [51]). These unusual characteristics are not obviously incompatible with action selection, but it is troubling that they have no obvious role in action selection either.

These features motivated Bar-Gad, Bergman and colleages [35, 37, 36] to propose that the main function of the basal ganglia is to reduce the dimensionality of diverse cortical data, and distribute it back to the cortex in compressed form. This compression would provide a means of making more information available to individual cortical neurons, since each cortical neuron can receive a limited number of synaptic inputs.

They further proposed that dopamine modulation of plasticity in the cortico-striatal synapses would bias this compression, so that information more relevant to rewards would be compressed with higher fidelity. They developed a computational model of dopamine-gated plasticity in cortico-striatal synapses, and showed that it caused high-fidelity compression of inputs that appeared in conjunction with dopamine signals, while other information was essentially ignored [37].

Plenz & Kitai [300] introduced a similar model of the striatum, including both dimensionality reduction and low correlations, although they did not address reinforcement bias.

2.3 Experimental Evidence

The action selection and reinforcement-driven dimensionality reduction (RDDR) hypotheses are different enough that one would expect little difficulty in differentiating them on the basis of experimental evidence. However, as discussed below, while different lines of evidence are more suggestive of one hypothesis or the other, most of the data can be reasonably interpreted as consistent with both. Nevertheless, 1) RDDR accounts more elegantly for certain anatomical and electrophysiological data, and 2) certain behavioural data support action selection. It will be argued in Section 2.4 that there may be no conflict between these observations, if RDDR is part of the mechanism for action selection.

2.3.1 Anatomical Evidence

Anatomical Funnelling

As discussed above, a major factor motivating the RDDR model is the funnel-like anatomy of the feedforward basal ganglia (BG) pathways. There are many fewer striatal neurons than cortical neurons projecting to the striatum, and many fewer

neurons in the output nuclei than the striatum. This reduction is consistent with the reduction in representational complexity hypothesized by RDDR. Furthermore, a two-stage reduction, with stronger nonlinearities in the first stage (consistent with strongly nonlinear responses of medium spiny neurons [280]) facilitates reduction of nonlinear input patterns [37], which may be important for compressing nonlinear relationships between contextual variables.

The cortico-striatal stage of funnelling is not particularly suprising in the context of action selection either. It just implies that the context for action selection is more complex than the action choices themselves (not necessarily than the parameters of action execution, control of which is managed by other structures). But the striato-pallidal reduction is more puzzling. Fewer neurons could imply either that 1) information is represented with lower fidelity in the output nuclei, or 2) less complex information is represented. But neither of these possibilities fits well with action-selection models.

For example, funnelling at this level might make sense if different groups of striatal neurons represented different candidate actions, while a single group of output neurons acted together to represent whichever action the striatum selected. However, the firing of movement-related neurons in the output nuclei varies with specific actions (e.g. movement at one joint in a certain direction), in much the same way as striatal neurons [13, 181]. Accordingly, action selection models assume parallel channels all the way through the direct pathway, each of which codes a different action independently. A more plausible alternative is that multiple independent modules in the striatum select actions independently, and these choices are somehow combined in the output nuclei [140]. This would account for a several-fold reduction in the number of neurons, but would not predict the observed 100-fold or greater reduction.

A more subtle possibility is that the output nuclei represent the same information as the striatum, with the same fidelity, but in a manner that permits fewer downstream computations. If n neurons are required to represent a value (e.g. an action) with a certain degree of fidelity, then roughly n^d neurons are required to support the computation of arbitrary nonlinear functions of d values [111].[2] If the output nuclei represent nonlinear functions of multiple action representations in the striatum, then the associated redundancy required in the striatum could easily explain its greater size. However, existing action-selection models have not proposed any such computations.

Finally, it must also be acknowledged that the neural code is not understood well enough that a tight correlation between numbers of neurons and representational complexity can be taken for granted. Interestingly, the average firing rates of output neurons are higher than those of the striatum. If spikes in each region convey similar amounts of information, then the reduction in representational capacity may not be as great as the anatomy suggests.

[2]This assumes near-linear synaptic integration, an assumption that is discussed at length in Chapter 4.

In summary, anatomical funnelling in the direct pathway is clearly consistent with RDDR, and it is unclear to what extent this anatomy is consistent with action selection, because the possibilities remain largely unexplored.

Granularity of Parallel Channels

Action selection models (e.g. [169, 52]) assume that feedforward basal ganglia pathways are made up of parallel channels. Each channel is taken to correspond to an action. Lateral interactions between channels in the direct pathway, and feedforward interactions in other pathways, are assumed to mediate competition between candidate actions [133]. There is ample evidence that the feedforward pathways consist of multiple distinct channels, but it is not clear that there are enough channels to separately code the thousands of actions that an animal might automate through reinforcement learning.

Projections from sensorimotor, oculomotor, limbic, and associative cortical regions fall on largely distinct striatal regions [293]. The cortico-subthalamic projection is also similarly segregated [209], and this separation is maintained in most other basal ganglia nuclei [16][3], and in the projections of the output nuclei through the thalamus to cortical targets [195]. Thus the cortico-basal ganglia circuit consists of functionally-distinct parallel circuits. Within the sensorimotor circuit there are further topographical distinctions (related to different body parts) in the striatum, pallidum, STN, and thalamus [16, 240, 131]. There is also evidence that loops connected with arm-related areas of the premotor, supplementary motor, and primary motor cortices are distinct [348].

However, other anatomical data suggest a high degree of feedforward integration. The dendritic fields of GPi neurons are large and disk-shaped, and incoming striatal fibres run perpendicularly to these disk-shaped fields, intersecting many of them [123], which suggests that single pallidal cells may integrate data from large ensembles of striatal neurons [298]. Consistent with this view, fibres from small areas of the striatum spread through relatively large areas of the GPi [153].

Quantitative results indicate that each pallidal neuron receives synaptic contacts from only about 1000 striatal cells, or about 0.001% of the human striatum, and a single striatal cell innervates only about 25 different pallidal cells [392]. Available anatomical data do not reveal whether each of these 25 pallidal cells receives input from the same 1000 striatal cells (consistent with a high degree of channel segregation), or partially overlapping sets of striatal cells, or largely different sets (consistent with low segregation within a few large channels).

Correlations between the firing patterns of pairs of GPi cells seem to suggest the latter. Cross-correlograms for pairs of GPi cells might be expected to show peaks around zero offset, for pairs of cells that receive many common inputs. Such

[3] Although motor and associative information is mingled in the SNr [293], and there is probably substantial collateral excitation within STN [134].

correlations are normally very rare [277, 306], suggesting that pairs of GPi cells receive input from weakly overlapping sets of striatal cells. However, low correlations are also predicted by RDDR, for neurons that share many inputs [37], due to differences in synaptic weights.

Interestingly, correlations increase with striatal dopamine depletion [277, 306, 51]. This has been taken as evidence that different pallidal cells normally represent information from different sources, but that this segregation breaks down with dopamine loss [51, 383]. These results can be interpreted as implying a lack of fine anatomical channel segregation.

This interpretation should be treated with caution, because correlated activity occurs mostly in conjunction with widespread firing rate oscillations [226]. However, dopamine-related changes in the specificity of movement-related activity support the same interpretation. While pallidal neurons normally respond in very specific conditions, for example in conjunction with movement in a certain direction at a certain joint [267, 118, 148], striatal dopamine influences this specificity. Responses to passive limb movement are both more common and less specific in GPi with striatal dopamine loss [118, 54]. Responses are also elicited by striatal stimulation in a greater variety of locations [366]. It has been proposed that dopamine regulates functional coupling between parallel subcircuits by vetoing divergent glutamatergic input to striatal neurons [51]. The effect may also relate to the modulation by dopamine of lateral inhibition in the striatum [144], or to the facilitation by dopamine of the firing of striatal projection neurons that are in a state such that firing is already likely [154].

In summary, 1) anatomical data do not specifically support the existence of fine-grained channels, 2) the apparent dopamine dependence of neurons' response specificity argues against anatomical segregation, if not against functional segregation, and 3) uncorrelated firing suggests that if small channels exist, they do not encode actions using correlated firing-rate increases or spike timing.

However, although fine-grained channels are often seen as a prediction or assumption of action selection models, their relevance to the hypothesis is questionable. Many of the same computations could, in theory, be carried out by either segregated or mixed channels. In particular, a large group of neurons that codes a d-dimensional vector can perform similarly to d separate groups of neurons that encode scalars. There are subtly-different implications for function approximation and robustness to noise in these two cases (discussed further in Chapters 3 and 8). But a distributed vector code would by no means preclude either independent activation of different functional channels, or competition between them. This would be true even if each neuron participated simultaneously in many different channels.

Convergence of Diverse Cortical Signals

Whether or not there any many fine-grained parallel channels through the basal ganglia, the segregation of coarse channels (particularly sensori-motor, cognitive,

and limbic) is clear. Coarse segregation raises key questions related to the action selection hypothesis: 1) how much action-relevant contextual information reaches motor cortex through basal ganglia pathways, and 2) does this contextual information converge with motor signals in a manner that is consistent with action selection? Action selection would be facilitated by convergence of motor and other contextual information in the striatum. Convergence at this level would allow diverse contextual information to drive an action channel. Importantly, it would also allow dopamine modulation of context-action mapping at corticostriatal synapses, consistent with the actor-critic analogy.

Axons from small regions of cortex diverge to broad areas of the striatum, so the parts of striatum innervated by non-adjacent points in the cortex necessarily overlap [403]. On smaller scales, this overlap often takes the form of inter-digitation of projections from different regions rather than convergence [335], although projections from functionally-related cortical areas sometimes converge [120]. Close interdigitation of cortico-striatal projections implies that even where their axons do not overlap, the dendritic arbor of a single striatal neuron may intersect axons from multiple interdigitated cortico-striatal projections, potentially integrating information from multiple sources.

However, as discussed above, the associative, motor, and limbic cortico-striatal projections, and large subdivisions within them, are largely separate. Furthermore, the associated cortico-BG loops appear to be largely closed on themselves. One line of evidence in this direction comes from a series of studies by Strick and colleagues. In early studies, they injected anterograde tracers into different motor cortical areas, and found little overlap between the labelled striatal regions, supporting the concept of distinct input channels. In further studies they injected Herpes virus into different cortical areas. Injection of a Herpes strain that neurons transport in the retrograde direction showed that different cortical areas receive input primarily from segregated areas of GPi [260], supporting the concept of segregated output channels. Furthermore, patterns of connectivity within the basal ganglia suggest that the input channel associated with a certain cortical area is connected to the output channel associated with the same area [348]. Similarly, later studies employed the rabies virus, which is transported retrogradely to first, second, and third-order neurons. Injections into different cortical areas labelled distinct third-order regions in the striatum. Injection of a conventional anterograde tracer in the primary motor cortex showed that this area projected to the same part of the striatum from which it received third-order projections [195]. The general impression from these studies is that the parallel loops through the basal ganglia are largely closed, i.e. that a region of basal ganglia that receives input from a certain cortical area will project back largely to the same area. However, these loops are not exclusively closed, because the different regions of the striatum overlap at their boundaries [192].[4]

[4]There is at least one other notable exception to the closed-loop rule [195]. Specifically, 10-20% of the third-order neurons in striatum that were labelled retrogradely from M1 injection were in the ventral striatum, a limbic region. This suggests a route by which limbic information can

A recent study used diffusion-weighted imaging tractography to confirm the presence of closed loops in individual human subjects [107]. This study also reported substantial overlap at the boundaries of the loops associated with different cortical areas. However, the latter result is difficult to interpret, because 1) sparse and dense connections were not differentiated, and 2) it was based on data that were merged across subjects (there may be individual variations across individuals on the order of scan resolution; for example, compare the two subjects in Figure 12 of [215]).

In any case, although there is some degree of overlap between the loops, the picture that emerges from these studies is not one in which each action channel receives all of the contextual information that might be relevant to its selection. Instead, it appears that the basal ganglia neurons that influence movement receive the *majority* of their information from sensori-motor cortical areas that code similar movement. A closed sensori-motor loop is perfectly consistent with proposals (e.g. [262, 59]) that multiple candidate actions are coded in the cortex and vetted by basal ganglia. But it is inconsistent with the idea that diverse contextual information contributes to this vetting [165, 350].

Of course, sensori-motor information is itself an important part of the context for future actions. For example, when a person is seated, proprioceptive data alone would be sufficient to turn off the reactions that maintain standing balance – a subtlety that is often absent in Parkinson's disease. Furthermore, projections from visual areas converge onto the sensory-motor striatum [43], providing a possible route for mapping visual stimuli to motor actions. However, information in the prefrontal cortex (e.g. working memory, declarative memory, and goal states) also bears on action selection. Closed loops do not argue against the action selection hypothesis, but they restrict it, by implying a relatively narrow range of selection-rule antecedents.

However, while cortico-striatal convergence consistent with the actor-critic model is in doubt, diverse contextual information certainly converges elsewhere in these circuits, specifically in the substantia nigra, thalamus, and cortex.

Midbrain dopamine neurons (primarily in SNc) project densely to the striatum, and receive reciprocal projections from the same striatal areas. However, projections from SNc to striatum terminate more broadly than the reciprocal projections. Furthermore, this greater breadth is directed in a hierarchy from limbic to cognitive to motor areas of the striatum [145]. This connectivity pattern suggests a hierarchy of reinforcement, in which affective information subserves reinforcement of cognitive performance, and cognitive information in turn subserves reinforcement of motor performance. However, the activity of dopamine neurons is spatially homogenous [328], so it is not clear to what extent this anatomical spiral corresponds to a spiral of information. In any case, this pathway is primarily modulatory, rather than excitatory (although some dopamine neurons co-transmit glutamate [72], so it is

affect cortical motor activity. Although the route from the ventral striatum to M1 is not clear, convergence with the motor channel is probably downstream of the striatum.

not likely to provide the type of convergence required by actor-critic-like selection models (e.g. [165, 350]).

However, a spiral of excitatory connections exists at the other end of the cortico-BG circuit, between the thalamus and the cortex. Cognitive and motor areas of the cortex are each connected reciprocally with different thalamic areas, but there are also non-reciprocal connections from limbic cortex to cognitive thalamus, and from cognitive cortex to motor thalamus [250]. Afferents from caudal motor areas also diverge to the medio-dorsal thalamic nucleus [173], which has strong reciprocal connections with the dorso-lateral prefrontal cortex. The thalamo-cortical loops therefore provide a pathway for communication across parallel channels.

Furthermore, this pathway may bring diverse context signals together with action signals in the striatum. The centromedian and parafascicular nuclei of the thalamus, which are the primary source of thalamostriatal neurons [343], do not participate in the spiral discussed above. Most studies to date have indicated that the participating thalamic nuclei (the ventral lateral, ventral anterior, and medial dorsal nuclei) project minimally to the striatum [188]. However, there is some recent evidence of more substantial projections (particularly from the ventral lateral nucleus [249]), which arise largely from thalamocortical collaterals [343]. So while the thalamo-cortical loops provide a site of cross-channel integration at the basal ganglia output, this integration may also extend back to the striatum.

Finally, cognitive and motor areas of the cortex are themselves directly connected [363]. An under-explored question is the extent to which the basal ganglia gate communication between different cortical circuits, rather than carrying information between them directly. Gating of cortical circuits would bring rich contextual data to bear on action selection, through the relatively massive intracortical connections, as compared with the low-capacity output of basal ganglia.

In summary, while the main pathways of the associative, limbic, and motor loops are largely closed, there are also several points of contact. Overlap at the boundaries of striatal regions provides a potential site of integration that would be consistent with the convergence of context and motor signals in the cortico-striatal projection. However, this convergence would be unavailable to action channels that originate away from these boundaries. In contrast, sites of more extensive integration (i.e. cortex and thalamus) may support more extensive use of converging information in selection, although it is less clear how these pathways relate to actor-critic-like reinforcement learning.

Lateral Inhibitory Interactions

Medium spiny neurons, the projection neurons of the striatum, inhibit both the internal and external segments of the globus pallidus. However, they also collateralize extensively within the striatum [387], inhibiting primarily other neurons of the same type. Action selection models that include these mutually-inhibitory in-

teractions (e.g. [165, 41, 384]) interpret them as a means of competition between conflicting action channels.

In the RDDR framework, they are instead taken to decorrelate the activity of neighbouring neurons, so that each neuron learns to extract different features of the input. This interpretation comes from theoretical studies of dimensionality reduction networks (e.g. [121, 122, 212]), which rely on lateral inhibition to prevent each neuron that receives the same input from redundantly extracting the same features (contrast with [281]).

These different interpretations lead to different predictions about the strength of lateral inhibition. Competition between actions would require strong enough synaptic interactions that the neurons at the head of a single action channel could silence neurons in other channels. In contrast, lateral inhibition in dimensionality-reduction models becomes weaker with learning, as the firing of neighbouring neurons becomes less correlated. So RDDR predicts weak lateral inhibitory synapses in mature networks.

In the first study to directly investigate the issue, Jaeger et al. [182] recorded intracellularly from striatal neurons while triggering action potentials in other nearby neurons, and failed to find any evidence of functional inhibitory interactions (despite dense physical connections). However, subsequent studies that averaged the membrane potential over many stimuli [85, 369, 210, 144] revealed weak, generally asymmetric interactions between a minority of neuron pairs. On the basis of these studies, Tepper et al. [357] argued that the lateral connections between medium spiny neurons were too weak to prevent or strongly modify their firing. Blocking inhibitory GABA receptors in the striatum can increase the firing rates of medium spiny neurons substantially [278], but Tepper et al. [357] attributed this effect to a small population of inhibitory interneurons. These interneurons are part of the feedforward cortico-striatal path, and therefore do not support lateral competition. Compared with lateral synapses, feedforward inhibitory synapses are about six times as strong individually, and twice as reliable. Furthermore, the feedforward neurons fire an order of magnitude faster. On the other hand, there are about two orders of magnitude more lateral than feedforward synapses onto an average medium spiny cell. Taking all of these differences into account, lateral and feedforward inhibition of a typical spiny neuron may have comparable magnitudes, when averaged over long enough periods of time. What is less clear is whether lateral synapses provide near-constant background inhibition that affects firing activity subtly (as argued by Tepper et al.), or whether small pairwise correlations in the activity of neighbouring medium spiny neurons add up across a large population, to produce stronger, more transient effects.

Tepper et al. [357] also took the fact that only a small proportion (10-15%) of the possible reciprocal connections between medium spiny neurons exist as an argument against RDDR. They pointed out that this sparse connectivity resembled that of a mature RDDR network in which weak synapses had been pruned, but that substantial new learning in such a network would require the formation of new

synaptic contacts, a process for which there is no direct evidence in the striatum. However, as shown in Chapter 8, there is no difficulty in performing dimensionality reduction in a network with sparse lateral connections. Sparse lateral connections result in stronger compression, with greater redundancy in the compressed code, which allows more flexible computations on the basis of compressed information.

In summary, the sparseness of lateral inhibitory connections does not rule out RDDR, and their weakness does not rule out competition between action channels. Furthermore, although some action-selection models have assumed that the dense collaterals of medium spiny neurons mediate competition, this assumption is not essential to the hypothesis. Other models (e.g. [262]) instead emphasize feedforward competition in projections between basal ganglia nuclei (although this mechanism would not permit disinhibition of the winning channel). Further complicating the situation, both excitatory and inhibitory terminals within the striatum are inhibited presynaptically [279, 214]. Depending on the fine structure of these connections, presynaptic inhibition might mediate negative feedback (for which there is evidence [278]), competition between feedforward pathways, or both. So, even if it is more clearly established in the future that lateral striatal connections are too weak to mediate competition, this will not rule out the action selection hypothesis. However, it will leave the hypothesis without an explanation for another prominent and unusual structural feature of the network.

Indirect Pathway, Hyperdirect Pathway, and Loops

Although the Albin/DeLong indirect pathway through the STN has been questioned, GPe neurons project directly to GPi [294], creating another, slightly shorter indirect pathway. STN axons (part of the Albin/DeLong indirect pathway) diverge extensively [326], but individual GPe axons form large multi-synaptic baskets around the cell body of a single target cell in GPi [294], synapsing only weakly onto other cells.

This pathway may supress actions that lead to poor outcomes. In support of this view, the D2-containing neurons, which are concentrated at the head of this pathway, are involved in learning to avoid bad decisions [125]. Remarkably, the vast majority of input from the motor cortex to the indirect pathway arises from collaterals of descending motor neurons [219], raising the possibility that motor actions are suppressed through this pathway largely on the basis of in-progress actions.

If the indirect pathway avoids poor outcomes, this could be taken to provide a separate argument against competition *via* lateral inhibition in the striatum. Competion between *selection* channels makes sense, because it would prevent attempts to perform two incompatible actions at once. But it would be counterproductive for an action-selection channel to compete with another channel that *supresses* the selection of a competing action. Specificity in the structure of lateral striatal connections would be needed in order to avoid this (e.g. direct pathway channels

should only inhibit other direct pathway channels). However, direct and indirect pathway neurons synapse onto each other [400, 28].

The cortex also projects directly to the STN, forming a "hyperdirect pathway" which bypasses the striatum, and through which cortical information reaches the output nuclei more quickly than through the direct pathway [272]. Frank *et al.* [126] provide evidence that this pathway signals the degree of conflict between action choices, and slows decision making in high-conflict situations, in people with advanced Parkinson's disease. STN lesion in an otherwise healthy brain (e.g. as a result of a stroke) can result in ballism, a severe condition that involves involuntary flailing movement [98], suggesting that the STN act routinely to supress impulsive movements.

Bar-Gad et al. [37] suggest that the hyperdirect pathway, which is less segregated than the direct pathway, might support more globally-aware compression (in parallel with localized compression in the direct pathway). Regarding the indirect pathway, if D2-containing neurons learn from pauses in dopamine firing, much like D1-containing neurons learn from dopamine bursts, this would suggest that as the direct pathway compresses context associated with reward (i.e. "exploitable context"), the indirect pathway might compress context associated with poor outcomes (i.e. "risky context"). From this perspective, relatively broad excitation of the output nuclei by STN would reduce represention of exploitable context and accentuate representation of risky context. This might (for example) draw attention to minor negative factors in high-conflict win-win decisions.

Summary

Although the action-selection hypothesis has received a great deal of attention, it does not yet account for two striking features of basal ganglia anatomy, i.e. funnelling along the main axis, and extensive lateral connections in the striatum. There is also insufficient evidence for fine channel segregation, a common assumption of action-selection models, but this assumption is not essential. Motor signals converge with all potentially-relevant contextual signals in the cortex and thalamus, but it is not clear whether there is enough convergence in the striatum to justify actor-critic models that have been advanced to explain reinforcement learning of actions. In contrast, RDDR is consistent with funnelling, lateral inhibition in the striatum, lack of channel segregation, and restricted convergence in the striatum. Finally, the roles of the indirect and hyper-direct pathways within the RDDR framework have received little attention.

2.3.2 Electrophysiological Evidence

Action-Related Activity

The striatum contains neurons that fire at a variety of phases of motor tasks, including preparation for movement, around movement onset, during the movement

itself, and while waiting for a cue to move [331]. In sequences of repeated movements, some motor striatal neurons fire at the onset of the sequence while others fire at the onset of each repeated movement [201].

The firing of many striatal neurons is related to the direction of movement [15], and some also fire in relation to force required for movement [82]. When target direction is dissociated from limb movement direction (using a computer display), most direction-related cells associate with target direction [14].

Downstream from the striatum, movement-related modulation of the firing of individual pallidal neurons also tends to correspond to movement of a certain joint in a certain direction [95, 267]. The firing of different cells in the STN is associated with movements of the limbs [97, 382] or eyes [115].

Broadly speaking, this movement-related activity is consistent with the selection of motor actions. However, while some striatal cells fire only with active movement, others fire regardless of whether movement is active or passive (i.e. induced by the experimenter), and some fire only with passive movement [18]. Responses to passive movement are also reported in the pallidum [97, 148], and STN [97, 382]. Activity related to passive movement obviously does not drive action selection. However, this activity is consistent with RDDR, since the actual movement of the body (regardless of the cause) is useful information, particularly for the planning subsequent movements.

Another important detail is the timing of movement-related activity. The activity of some BG neurons is modulated before the earliest myoelectric activity, while that of the majority is modulated later [12, 181]. The same can be said of neurons in the cortex and cerebellum [98]. However, BG activity tends to change after cortical activity (e.g. [128, 237]). For example, although some striatal cells fire well before movement onset, cortical cells in the supplementary motor area fire still earlier [15, 317]. Furthermore, movement-related firing activity in the basal ganglia often persists well beyond the completion of the associated movement [181].

Perhaps more strangely, Lau & Glimcher [215] report a subset of neurons in the primate striatum with activity tuned to movement direction, but beginning substantially after movement *completion.* This activity occurred in a reward-delivery period of the task, but was not sensitive to actual reward delivery.

All of this timing is consistent with RDDR, but reconciling late movement-related activity with action selection is more complex. As discussed above, some action selection models propose that candidate actions are vetted in the basal ganglia only after they are initially coded in the cortex. This would account for the tendency for movement-related activity to occur later in basal ganglia than cortex, but not for the fact that much movement-related basal ganglia activity follows movement initiation [265, 26], and that some [215] follows movement termination. Another possibility is that automated actions that are obtained by reinforcement learning in the basal ganglia are subsequently transferred to the cortex. Well-learned actions might then be selected in parallel by both the cortex and basal ganglia, but the cortex might be slightly faster.

A recent study by Kobayashi et al. [206] provides more specific support for the reinforcement-driven bias in striatal coding that is hypothesized by RDDR. In this study, primates performed saccades to cued locations. Rewards were delivered with saccades in a single direction only, and the direction varied in different trial blocks. Neural activity in the cortex varied with the required saccade direction. However, the activity of many neurons in the caudate nucleus varied with saccade direction only when the saccade direction corresponded to the rewarded direction.

In summary, responses to passive movement, the late timing of movement-related activity, and reinforcement-contingent coding of some movements are all consistent with RDDR. But in the context of action selection, they imply 1) other redundant (and faster) selection mechanisms outside basal ganglia, and 2) non-selection-related activity within basal ganglia.

Reward-Related Activity

A series of studies by Schultz and colleagues (reviewed by [329]) showed that the activity of midbrain dopamine neurons reflects errors in the animal's prediction of rewards. This pattern of dopamine neuron activity has since been confirmed by several other labs (reviewed by [330]). As discussed in Section 2.2, this activity corresponds to the reward-prediction error signal of TD learning.

Dopamine-mediated reward prediction error figures prominently in action selection models based on the actor-critic architecture (e.g. [165]). These models are appealing because 1) dopamine modulates corticostriatal plasticity, consistent with the proposed role of the critic in modifying context-contingent choices of the actor, and 2) as discussed above, TD learning suggests a powerful means by which an animal can develop a behavioural policy that maximizes rewards. Notably, a TD-like mechanism would allow an animal to associate a reward not only with the decision that immediately precedes it, but also with a chain of earlier decisions that may also have been critical for obtaining the reward.

In support of this interpretation, abstract TD-learning models predict the behaviour of human subjects in a variety of reinforcement learning tasks, and reward predictions and errors in TD models correlate with various fMRI and EEG signals (reviewed by [74]). Despite these parallels, detailed basal ganglia models based on the actor-critic architecture have been criticised as making physiologically-unrealistic assumptions, and of not modelling the actor component in enough detail [185]. However, alternative architectures have been proposed, which agree more closely with biology, while supporting essentially the same role for dopamine in reinforcement-based action selection (e.g. [93, 289]).

More fundamental challenges to the critic analogy arise from the behaviour of dopamine neurons in non-rewarding situations. Dopamine neurons typically exhibit a pause in firing after aversive events such as a pinch, aversive air puff, etc., which is similar to their response to reward omission. However, a minority of neurons in the dopaminergic nuclei increase their firing in response to these stimuli, and

the neurons that pause sometimes fire more quickly after pausing [330]. Positive responses to aversive stimuli are at odds with the general pattern of agreement between dopamine activity and reward prediction error, so they present a potential challenge to reinforcement-learning models based on this view of dopamine. However, there is evidence that the neurons with short-latency positive responses to aversive stimuli belong to the small minority of non-dopaminergic neurons in these nuclei [370].

Dopamine neurons also burst briefly and with short latency to novel but apparently neutral sensory stimuli [370, 308]. These responses can be reconciled with reinforcement learning in a number of ways. For example, some visual stimuli may be inherently rewarding (as argued by [53]). Also, a dopamine signal that conflates reward with novelty might encourage exploratory behaviour [370].

In summary, despite ongoing challenges, the activity of dopamine neurons is essentially consistent with the role of the critic, in reinforcement learning of an action-selection policy. In contrast, the RDDR hypothesis is not sensitive to the precise pattern of dopamine activity. Generally, its correlation with reward prediction error would bias the compression process toward the earliest contextual information from which rewards could be predicted, but there would be little harm (and possibly some benefit) in also compressing novel stimuli with higher fidelity.

Stimulus-Related Activity

Somatosensory, auditory, and extrastriate visual cortex project substantially to the striatum. These sensory signals could potentially provide useful contextual information, which reinforcement learning processes could map onto striatal action representations. For example, a particular visual stimulus could be mapped onto an appropriate motor response.

But striatal neurons also respond directly to sensory stimuli. For example, in the somatosensory domain, many neurons in the putamen respond at short latency (25-50ms) to loads that are applied externally to the arm [83]. Whether these responses are really sensory is not beyond doubt, because load application triggers transcortical reflexive motor responses with EMG onset at slightly greater latency [245] (so this activity could be driven either by somatosensory neurons, or by collaterals of the same motor neurons that drive EMG activity after a descending conduction delay). However, tactile stimuli also elicit striatal activity [60], and Parkinson's disease impairs performance in sensory discrimination tasks [60].

Visual stimuli also elicit responses in the striatum. Some of these responses might actually reflect reward prediction errors, in cases where the stimuli are associated with an opportunity for the animal to earn a reward (e.g. [12]). Others are enhanced if the animal makes a saccade to the stimulus [155]. However, while some nominally-sensory responses in the striatum could be interpreted as motor or reward signals, others are less ambiguous [60], and are therefore inconsistent with a narrow view of action selection.

On the other hand, the basal ganglia probably also influence sensory cortical processes. They project (narrowly) to the visual cortex [259]. Functionally, lesions impair vision in infants [257], and visual hallucinations are a common side-effect of dopamine medication in Parkinson's disease [117]. It is therefore possible that the basal ganglia influence sensory processes in much the same way that they influence motor and cognitive processes (for example, selectively).

Summary

To summarize, 1) the activity of dopamine neurons is consistent with either action selection or dimensionality reduction; 2) sensory responses in the striatum argue against a purely motor/cognitive selection function, but they are consistent with a selection-related sensory function; and 3) in contrast, movement-related activity (particularly responses to passive movements and the late timing of basal ganglia activity) are more difficult to reconcile with action selection than with RDDR.

2.3.3 Behavioural Evidence

The action-selection hypothesis makes fairly clear predictions about the influence of the basal ganglia on behaviour. In particular, experimental manipulations of basal ganglia should show clear roles in driving movement, and in distinguishing between alternatives on the basis of past reinforcement. This section argues that for the most part the experimental evidence bears these predictions out.

In contrast, novel behavioural predictions of RDDR (as opposed to electrophysiological predictions) have been slow to emerge. This is not surprising, because the effects of RDDR on the cortex would be more subtle than those of action selection. On the other hand, since the basal ganglia project densely to cortical motor areas [109], one might ask what aspects of motor control would benefit particularly from the reinforcement-biased, compact representation of context that RDDR would provide. Different degrees of involvement might be expected in different motor behaviours, but this type of input would seem to be more useful for high-level choices than for low-level control. In other words, the basal ganglia might affect similar types of overt behaviour regardless of whether they perform RDDR or selection, although RDDR should have a more subtle influence on a broader range of behaviour.

Involvement in Movement Execution

As suggested indirectly by disease symptoms, activity in the basal ganglia does not just reflect movement, but can cause movement as well. Microstimulation of the motor striatum can produce muscle activity at a latency of about 20 ms [201]. It can also produce movements (largely about a single joint) that are scaled with stimulation strength [17, 18]. These data are compatible with RDDR, in

that stimulation might approximate the representation of a contextual trigger for movement. They are also clearly consistent with action selection.

However, the basal ganglia are also implicated in aspects of movement execution that are less-clearly related to selection. For example, Parkinson's disease is associated with movements of reduced amplitude, including small strides [270] and handwriting [284], suggesting basal ganglia involvement in movement scaling.

The symptoms of Parkinson's disease should be interpreted with caution, because 1) the disease also damages neurons outside the basal ganglia [55], and 2) gradual onset of symptoms gives the rest of the brain ample opportunity to adapt, possibly leading to secondary effects.

However, the influence of the healthy basal ganglia on movement has been studied more directly, through acute manipulations of the output nuclei. Studies in monkeys, in which the GPi was cooled [163], its activity inhibited by the GABA-receptor agonist muscimol [266, 175, 378], or its cells lesioned with kianic acid [162, 266], have reported slowed arm movements and reduced muscle activity, usually with no effect on reaction times. Some studies also reported excessive co-contraction of antagonist muscles [266], or variable reaction-time effects [175]. These variations may have related either to the involvement of the GPe, and/or involvement of associative as well as motor loops. One study [194] reported co-contraction with GPe lesion, but not GPi lesion.

A recent study that carefully isolated injections to motor areas of the GPi reported reduced muscle activity, and slowed movements that undershot their targets [100]. These lesions had little effect on reaction times, and did not cause co-contraction [100].

Together, these results suggest that the role of the healthy basal ganglia in motor control is not limited to selection, but extends to movement parameterization.[5]

Involvement in Reinforcement Learning

In a well-known study, Knowlton et al. [204] showed a double dissociation between 1) declarative memory, and 2) implicit memory for probabilistic stimulus-outcome contingencies, which implicated the basal ganglia in the latter. Subjects performed a "weather prediction" task, in which they tried to predict the outcome "rain" or "shine", after viewing sets of neutral images with which these outcomes were probabilistically associated. Patients with Parkinson's disease were unable to improve their performance with practice. In contrast, patients with amnesia did improve with practice, but had great difficulty answering follow-up questions about details of the experiment. These results provide evidence that the basal ganglia are involved in probabilistic classification based on trial and error experience. In this

[5]It has also been suggested that the basal ganglia contribute to on-line motor control through selection of discrete corrective sub-movements [166]. However, related pallidal activity may not be early enough to drive these corrections (see Figure 8 in [166]), and on-line corrections are unimpaired by GPi lesions [100].

early study, the difference in classification performance between the Parkinson's and amnesic groups was not actually statistically significant, and disappeared after extended practice. However, subsequent studies provide additional support for the role of the basal ganglia in probabilistic reinforcement learning [304, 150, 338].

In the classification task of Knowlton et al. [204], participants had two ways to improve: 1) by repeating choices that led to success in the past, and 2) by avoiding choices that led to failure. Frank et al. [127] dissociated these two mechanisms, and found evidence for striatal involvement in both. Subjects were repeatedly shown three pairs of visual stimuli (Japanese characters), asked to choose one stimulus from the pair, and then told whether or not they chose correctly. The correct choices were determined randomly with different probabilities. For one pair (which the authors called A and B), stimulus A was the correct choice 80% of the time. In other pairs the probability of each stimulus being correct was closer to 50%. After a training phase with feedback, A and B were each paired with more neutral stimuli, in a testing phase without feedback. This allowed separate assessment of the participants' tendency to choose A as opposed to avoiding B. Subjects with Parkinson's disease excelled at learning to choose A while on dopamine medication, and excelled at learning to avoid B while off medication. Intriguingly, patients off medication appeared to be better than age-matched controls at learning to avoid B. These results implicate the basal ganglia and dopamine in both "Go" and "NoGo" learning based on reinforcement.

What are the physiological correlates of Go vs. NoGo learning? As discussed previously, dopaminergic neurons burst when rewards are higher than expected, and pause when rewards are lower than expected [330]. Furthermore, dopamine modulates plasticity at corticostriatal synapses [312]. One possibility is that dopamine bursts strengthen excitatory cortical synapses onto the D1-containing neurons at the head of the direct pathway, and that dopamine pauses strengthen the corresponding synapses at the head of the indirect pathway. A later study by Frank et al. [125] showed variations in Go/NoGo learning in healthy subjects with genetic differences related to striatal D1 and D2 receptors, supporting this interpretation.

These results implicate striatal D1 and D2 receptors in learning to select actions with good outcomes and to avoid actions with bad outcomes, respectively, consistent with the hypothesis of action selection via reinforcement learning.

Reconciling these results with RDDR is less straightforward. There are several possible interpretations. Firstly, in successful and failed trials, respectively, the direct and indirect pathways might code whichever stimuli are presented (e.g. A and B). In this case, RDDR might encourage recognition of these stimuli by cortical circuits, but it would provide no basis for distinguishing between them. This possibility conflicts with evidence of striatal dopamine involvement in differentiating these stimuli. Secondly, RDDR might preferentially code the stimulus to which the subject directs the most attention in each trial, which is likely to be the one that the subject chooses. In this case, RDDR would provide a basis for distinguishing between stimuli. For example, A would be coded preferentially by the direct

pathway, as part of the context of success, and B would be coded preferentially by the indirect pathway, as part of the context of failure. This possibility might be tested by forcing subjects to attend to the non-selected stimulus prior to feedback. Finally, RDDR might compress cortical representations of actions (e.g. "select A").

The latter two cases could operate simultaneously, and would be consistent with basal ganglia activity related to sensory input, and the late timing of movement-related activity, respectively. In either case, RDDR would not be an alternative to action selection, but part of its mechanism. In particular, compression of contextual information might be one step in the process of reinforcement-guided selection.

Involvement in Habits

Yin et al. [395] showed that stimulus-response habits in rats (as opposed to goal-oriented actions) rely on the dorsolateral striatum. Rats were trained extensively to press a lever for sucrose rewards. Sucrose was then devalued in some rats by pairing it repeatedly with lithium chloride injection, which induces nausea, so that the rats no longer consumed sucrose when they had free access to it. When exposed again to the lever, intact rats continued to press it, despite devaluation of the sucrose outcome. However, rats with lesions of the dorsolateral striatum pressed the lever much less often when sucrose was devalued. In a separate study [396], rats with similar initial training were (later in the experiment) rewarded instead when they refrained from pressing the lever. Rats with muscimol infusion in dorsolateral striatum reduced lever pressing when the reward contingency changed, but control rats did not (these differences persisted the following day without drug infusion, so this was not due to a simple impairment in the ability to press the lever).

These results directly implicate the dorsolateral striatum in habitual stimulus-response action selection. In order to interpret these results within the RDDR framework, it must be hypothesized that these habits are formed primarily on the basis of compressed contextual data. The establishment of compressed striatal representations in habit learning might explain why as habit learning progresses, fewer striatal neurons are active in association with habitual movements, but those few are active more strongly [355], and the activity of striatal units increasingly focuses around specific phases of a more complex task [40].

2.3.4 Summary

The behavioural predictions of RDDR are less definite than those of action selection. However, RDDR might be expected to impact movement more broadly and subtly than a selection mechanism. The influence of the basal ganglia on movement scaling is consistent with this interpretation.

The involvement of the basal ganglia in reinforcement learning tasks and habitual actions is clearly consistent with the action-selection hypothesis. If the basal

ganglia perform RDDR, rather than selection, this involvement implies that information compression is important for these behaviours. Section 4 discusses this possibility in more detail.

2.4 Compatible Theories

2.4.1 Series Hypothesis

Anatomical and electrophysiological evidence are consistent for the most part with both the action selection and RDDR hypotheses. However, RDDR provides a simpler explanation for several observations, including anatomical funnelling, weak (but dense) lateral inhibition in the striatum, late movement-related activity, and responses to passive movement. On the other hand, although the effects of GPi lesions suggest a role for the basal ganglia in movement parameterization, experiments in reinforcement learning implicate the basal ganglia motor loops in habitual action selection.

If alternative hypotheses are consistent with different lines of evidence, one wonders whether they are mutually exclusive. RDDR and action selection are compatible in series, in that actions could (in theory) be selected on the basis of compressed rather than raw contextual data. A series architecture would account for diverse experimental evidence.

Furthermore, compressed contextual data would facilitate selection, in that it would enable convergence of a greater amount of relevant contextual information onto individual selection neurons. It would also filter out reward-irrelevant context signals, which would facilitate learning by reducing noise arising from spurious context-action relationships. The information entering the selection system would therefore be organized in a manner that emphasized important environmental cues for action, based on the animal's accumulated experience.

More importantly, RDDR would facilitate generalization. When an environment is complicated and/or the context variables are continuous, an agent encounters novel situations routinely, and performance depends critically on generalization from previously-encountered states [353]. If action selection were based on compressed contextual data, then compression should make this generalization more robust, by reducing the effective dimension of the context space without reducing the number of samples.

Similarly, reinforcement learning requires estimation of the reinforcement value of each state. This process requires the same type of generalization across distinct contexts, and would benefit in the same way from appropriately-compressed contextual data. Sahani [320] discusses essentially the same issue from another perspective, pointing out that the reinforcement value of a context is likely to be more parsimoniously associated with underlying causes than with sensory data. He then proposes a biologically plausible mechanism by which diffuse reinforcement

signals might optimize sensory coding in the cortex, for use in value-function approximation. RDDR in the striatum would provide a complementary mechanism for achieving the same goal. Specifically, instead of optimizing sensory codes to support extraction of reward-relevant latent variables, the extraction of latent variables would be biased along reward-relevant dimensions.

Relatedly, Swinehart & Abbott [354] showed that dimensionality reduction facilitates reinforcement learning of function approximation (this is distinct from the problem of value-function approximation during reinforcement learning of a selection policy). In their method, a target function is approximated by weighted basis functions. The weights are changed randomly, and reinforcement consolidates changes that lead to improved approximation. With a large number of basis functions, the large dimension of the weight space slows this semi-random approach to the optimal weights. Compressing the input simplifies the search. At first glance, it seems that a similar mechanism might underlie action selection in the striatum. Specifically, an intrinsically random element in striatal neurons could select actions at random, and with success, reinforcement could strengthen weights associated with coincident contextual signals, making the successful action more likely to be performed again in the same context. However, closed loops and timing of action-related activity (discussed above) suggest that if the striatum does select actions, it does so by vetting candidate actions that are coded in the cortex rather generating its own actions at random. On the other hand, if the striatum instead represents compressed contextual data, reinforcement signals within the cortex might shape function approximation based on its output, in a similar manner.

2.4.2 Site of Context-Action Mapping

If RDDR and action selection operate in series, where is contextual information mapped to actions? One possibility is that the mapping occurs in the projection from the striatum to the output nuclei (so that the output nuclei represent action signals). However, responses to passive movement observed in the output nuclei, and the later timing of movement-related activity there compared with the cortex, argue against this. It is also unclear how mappings at this site would be established and maintained. As argued further in Chapter 8, either reinforcement or supervisory signals are needed in order to establish a mapping from one type of information to another (e.g. from contextual information to action; as opposed to simply reorganizing contextual information). Dopamine innervation of the output nuclei [381] might play a role, although it is not as dense as in the striatum. The hyper-direct pathway could conceivably provide a supervisory signal, but this role would be inconsistent with extensive collateralization and divergence in this pathway, and with evidence for its role in slowing decisions [126].

In contrast, thalamic or cortical mapping would allow convergence of RDDR output with prefrontal signals that subserve other, more cognitive modes of action selection (e.g. declarative memory of what has worked in the past; instruction following). This would provide a pathway for automation of these cognitive selection

mechanisms over time. Specifically, these cognitive signals would provide appropriate teaching signals for supervised learning of context-action mapping, so that intermediate cognitive steps could eventually be skipped.

The possibility of automating cognitive processes in this manner does not preclude the additional possibility that parallel processes might explore the action space in a more random manner. Both the cortex and the thalamus [344] receive substantial dopamine input, which could consolidate successful variations [213]. Accordingly, EEG signals that reflect reward prediction errors are found over the motor cortex on the side used to carry out a decision [75]. However, if relevant cognitive information were available, it would make exploration more efficient. This type of transfer from cognitive to automated control would essentially cause the animal to behave as it had in the past, unless it paid attention. This would be consistent with shifts in activity from motor to prefrontal circuits, when human subjects pay attention to their performance in overlearned motor tasks [189].

Finally, although context could plausibly be mapped onto actions in either the thalamus or cortex, it is not certain that the mapping occurs in one step, or that context and action signals are anatomically segregated.

2.4.3 Experimental Tests

Although there are several possible sites of context-action mapping, the basic question relevant to the RDDR hypothesis is whether this mapping occurs primarily in the cortico-striatal projections, i.e. whether striatal activity is more closely related to context or action (Figure 2.1). A basic strategy for addressing this question would be to 1) train a habitual response in rats, 2) block synaptic plasticity in the striatum, 3) remap the stimulus to a new response, and then 4) test whether the new mapping is sensitive to reward devaluation. If not, i.e. if the new response is habitual, this would suggest that the striatum had coded the relevant context for action, while the mapping to a selected action was performed elsewhere.

Such an experiment would resemble the study of Yin et al. [395], which provided evidence for the role of the dorso-lateral striatum in habitual actions (as discussed above). In their experiment, rats learned to press a lever to obtain a sucrose reward. Rewards were provided on a variable-interval schedule, in order to make the rats' responses habitual (response extinction is rapid when rewards are devalued after continuously-reinforced training; e.g. [114]). The interval schedule dissociates the lever-pressing response from the rats' expectation of reward, so that the antecedent for the lever-pressing action becomes simply the opportunity to press the lever. This experiment does not clarify whether the critical information coded in striatum is the availability of the lever, or the lever-pressing action itself.

In order to test this distinction in a similar experiment, rats could initially be trained to manipulate a lever in some way (e.g. push it to the left), and then trained to manipulate it differently (e.g. move it to the right) after blocking striatal plasticity. Long-term synaptic potentiation in the striatum can be inhibited by

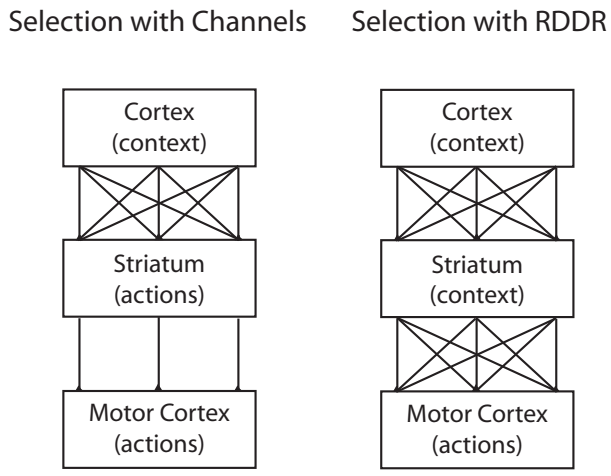

Figure 2.1: With respect to the possible role of RDDR in action selection, the fundamental question is whether context-action mapping occurs in projections onto the striatum, or later in the feedforward pathways.

NMDA-receptor blockade (e.g. [29]). Long-term depression can proceed independently of NMDA [238]. However, it requires retrograde signalling at the synapse, and is sensitive to mRNA translation inhibitors [393]. After inital training to establish stimulus-driven behaviour, striatal plasticity would be minimized by local drug injections, and the reward contingency changed (e.g. to rightward lever push). Only one action should be possible in the second phase, because alternative actions remain goal-directed even after extensive interval training [394]. If the contribution of the dorso-lateral striatum to habitual actions is to represent exploitable context, this representation should remain intact while plasticity is blocked, and newly learned responses to this contextual information should be resistant to reward devaluation.

2.5 Conclusion

The action selection and RDDR hypotheses are attempts to summarize the function of a very complex system. There are inherent limitations associated with this goal: 1) neither hypothesis predicts the complex anatomy and physiology of the basal ganglia in detail, and 2) each basic hypothesis can be interpreted in a number of ways (one consequence of which is the wide variety of published action selection models). Ultimately, it is doubtful whether the behaviour of these 10^8 coupled, nonlinear neurons can be reflected faithfully in a simple model. But the system can be understood more clearly if we can identify the simple model that is minimally misleading.

The anatomical and electrophysiological evidence reviewed here leans in favour

of RDDR, but the behavioural evidence supports action-selection. These potentially-conflicting interpretations can be reconciled by the further hypothesis that dimensionality reduction in the striatum serves as the input stage for action selection in the pallidum, thalamus, and/or cortex.

2.5.1 Reinforcement Learning

Dopamine responses resemble reward prediction error in the temporal-difference model of reinforcement learning, which makes it appealing to think of the striatum as the actor in the actor-critic model. The series hypothesis presented here contradicts this idea, and this is itself a reason for skepticism. Although detailed models comparing the actor-critic architecture to basal ganglia have limitations, the general outlines of the actor-critic concept of basal ganglia have received a great deal of attention and scrutiny (much more than RDDR).

Part of the appeal of the actor-critic concept of basal ganglia is that it goes a long way toward doing away with the homunculus. TD learning leads to complex, adaptive behaviour in models. Because the basal ganglia are conserved across many species, it is tempting to think that something like TD learning in the basal ganglia might account for much complex, adaptive animal behaviour.

The series hypothesis separates the reinforced selection policy from the part of the brain where it is expected, i.e. the dopamine-dense striatum. However, midbrain dopamine neurons also project throughout the cortex, and influence cortical plasticity [213]. So the series hypothesis does not argue against basal ganglia involvement in reinforcement learning, but suggests a way in which reinforcement learning could be made more practical for complex environments, i.e. by structuring context representation to emphasize salient features, and facilitating generalization to novel situations.

2.5.2 Next Steps

It makes little sense to discuss principles of basal ganglia function without considering what the basal ganglia do. Because of the complexity of the basal ganglia circuits, and the central location of the basal ganglia in the brain, their function is not obvious. However, the literature reviewed here suggest that 1) RDDR is consistent with striatal anatomy and physiology, 2) action selection is consistent with the influence of basal ganglia on behaviour. This chapter has further argued that RDDR may serve as an input stage for action selection. One way forward from these arguments would be to encapsulate them in a sophisticated computational model, and compare the model's behavior with as many observations as possible. However, a directly-relevant experimental test is possible, as outlined above, so that would be a more appropriate first step.

Furthermore, while a great deal is known about the anatomy and physiology of basal ganglia circuits, questions remain about their basic computational properties. For example, what computations can inhibitory projections perform? How do cell-intrinsic firing dynamics interact with computation? Although useful basal ganglia models have been developed in the past, without addressing questions like these, more sophisticated models will increasingly require answers. Because this is a theoretical thesis, the experimental work is deferred, and the remainder of the thesis focuses on elucidating the basic computational properties of the basal ganglia circuits. In conjunction with future experimental work, the computational properties explored in the following chapters should ultimately contribute to more advanced and comprehensive models of basal ganglia function.

Chapter 3

Population Coding

This chapter reviews theoretical aspects of population coding, i.e. information representation by distributed patterns of neural activity, which is a key feature of information processing in the basal ganglia.

Whenever an externally-identifiable signal (e.g. reward delivery, movement about a certain joint, etc.) is correlated with basal ganglia activity, it is correlated with not one, but many basal ganglia neurons. This is why electrophysiologists are consistently able to find neurons responsive to the particular conditions of their experiments, despite the fact that a typical experiment will examine less than .001% of the neurons in the basal ganglia. Furthermore, individual basal ganglia neurons sometimes participate in the coding of diverse information, e.g. movement direction in combination with visual stimuli [12, 146], applied load [83], or reward expectancy [27]. Thus the basal ganglia represent information in the form of distributed neural codes, or population codes. In this respect, the basal ganglia are typical of mammalian neural systems.

This chapter reviews population-coding, and introduces the Neural Engineering Framework (NEF), a coherent theory of population coding that also addresses the relationship between population codes and network dynamics. Finally, the chapter closes with a discussion of "cosine tuning", which plays an important role in the NEF.

3.1 Introduction

Not all neural systems use population codes. In some invertebrate circuits, each neuron has a distinct and well-defined role, which is conserved across individuals of a species, and circuit behavior can be strongly affected by damage to a single neuron. The abdominal ganglion of the snail Aplysia [191] is a well-studied example. Still other systems are hybrids. For example, in the cricket cercal system, about 1000 correlated primary sensory neurons project to a small number of identified interneurons that have distinct roles [359].

However, population coding is ubiquitous in mammalian central nervous systems. In these circuits, any single item of information is represented by many neurons, so that the activity of any single neuron has little impact on the representation of that information. For example, a visual stimulus moving across a certain part of the visual field, in a certain direction, will cause firing activity in a large number of neurons in the middle temporal area of the primate cortex [246]. Collectively, the firing of these neurons contains information about the motion of the visual stimulus (i.e. encodes the stimulus). Similarly, populations of neurons in the primary visual cortex code for the orientation of visual contours [168], and populations in the hippocampus code for spatial location, which can be inferred from multiple sensory modalities [282].

One reason population codes are useful is that the activity of any one neuron typically has an ambiguous relationship with the underlying signal. This is because individual neurons are noisy, and typically active over a wide range of signal values (wider than the resolution needed by the system). Studies by Georgopoulos et al. [129, 130] provided the first clear illustration that populations of such imprecisely-tuned neurons can carry precise information. They recorded the firing activity of neurons in the arm area of the primate motor cortex, and found that individual neurons fired most quickly during (and just prior to) movement of the arm in a certain direction, which they called the neuron's "preferred direction". However, neurons were broadly tuned: they also fired at above-baseline rates during movements in quite different directions, up to about 90 degrees from the preferred direction. The broad tuning of these neurons appeared to conflict with the precision of movements under their control. However, Georgopoulos et al. were able to accurately predict movement direction from the recorded activity of multiple neurons. To estimate movement direction from population activity, they used a weighted average of preferred direction vectors, with each vector weighted by the increase in its activity above baseline. This method provided reasonably accurate predictions of movement direction, even during delay periods prior to movment. Their success demonstrates that population activity as a whole contains accurate information about the monkey's intended movement direction, despite the broad tuning of individual cells. This principle is vividly confirmed in more recent experiments, in which restrained monkeys can feed themselves using robotic arms that are controlled by the decoded activity of neurons in the motor cortex [373].

Throughout the cortex, the activity of neurons in a small neighbourhood is typically correlated. Therefore, regardless of whether a corresponding external signal can be identified, it is reasonable to think of the population as coding *some* underlying signal – if not a sensory or motor signal, then perhaps a complex transformation of past and present sensory and/or motor signals.

Information coding through correlated neuronal activity has a clear disadvantage in terms of energy efficiency. The human brain uses about 20% of the energy of the whole body. Most of this energy could be conserved by reducing the redundancy inherent in population coding. Alternatively, doing away with population coding would increase the representational capacity of the brain by one to three

orders of magnitude. Not surprisingly, in the face of this pressure, population coding also provides important advantages. Three key advantages are discussed in the following sections.

3.1.1 Noise Reduction via Redundancy

There are many sources of noise in neural systems. For example, it is common for a presynaptic terminal to fail to release synaptic vesicles when a neuron spikes (e.g. [19]). A practical advantage of population coding is that if the noise in different neurons is statistically independent, then the accuracy with which a population represents information improves indefinitely with increasing population size (by the central limit theorem).

However, neuronal noise is often correlated, and correlated noise limits coding accuracy, regardless of the number of participating neurons. Zohary et al. [404] showed that weak covariation observed in the firing rates of pairs of cortical neurons (in the middle temporal area) has a large effect at the population level. Specifically, the predicted signal to noise ratio of the population code does not increase indefinitely with increasing population size, but saturates with about 100 neurons. This limits the useful size of a population for coding accuracy purposes (although noise reduction is not the only benefit of a population code, as discussed below).

Interestingly, noise correlations can also improve coding. As a simple example, if a value is encoded in the *difference* between two firing rates, then the code is not corrupted by random co-variation in these rates, whereas it is corrupted if the rates exhibit independent random variations. More generally, if signal-related and noise-related correlations between different neurons are in opposite directions, correlated noise corrupts a population code less than noise that is uncorrelated between neurons [31].

In the same way that a population code reduces noise under normal circumstances, it will also reduce the impact of damage to part of the population (this is clear if one thinks of damage as low-frequency noise). Furthermore, the fact that a population-coding network is relatively insensitive to damage to a small number of neurons may provide flexibility that is useful during learning.

3.1.2 Computation via Diversity

A second benefit of population coding is that, compared to a single-neuron code, a population code can support more sophisticated and flexible transformations of represented signals. The scope of transformations that can be performed in a projection from one population to another can be understood by analogy with artificial neural networks. A feedforward network with one hidden layer of sigmoidal units can approximate any smooth function with arbitrary precision, provided the hidden layer is sufficiently large [84]. This is intuitive for a network with one input

neuron. In such a network, the activity of each neuron in the hidden layer is a function of one variable, i.e. the firing rate of the input neuron. If different hidden-layer neurons have different weights and biases, their firing rates become infinitely varied functions of the input. If there are enough of them, they can span any finite-dimensional space of functions of the input. An output neuron can then decode any function of the input, through an appropriate linear combination of hidden-layer activities. Similarly, networks with n input neurons can approximate functions of n dimensions. This is true for any neuron model in which output is a non-linear function σ of the sum of inputs and an intrinsic bias, where σ approaches zero as the net input approaches negative infinity, and saturates as the net input approaches infinity [84]. Therefore the result applies not only to artificial neural networks, but also to a wide variety of more physiologically-realistic neuron models.

In the context of population coding, the firing rates of neurons in a population can be interpreted as functions of the low-dimensional information that they encode, rather than higher-dimensional functions of the firing rates of the many correlated neurons that drive them. Much like the hidden-layer neurons discussed above, if there are enough neurons in a population, and enough variety in their tuning curves, then arbitrary functions of the encoded space can be extracted by a post-synaptic neuron through appropriate choice of synaptic weights (Figure 3.1).

The diversity of tuning curves that is required for flexible transformation can arise simply from random variation between neurons. Further diversity may be enforced by competitive interactions between different neurons within the population. The pattern of diversity can also be optimized for the task, resulting in more efficient use of neurons. Such task-related focusing of tuning curve diversity is essentially what the back-propagation algorithm [319] achieves in hidden layers of a feed-forward network. The back-propagation algorithm itself is not physiologically realistic [81]. However, there are more realistic algorithms that perform comparably [390].

3.1.3 Representation of Uncertainty

Finally, population codes provide a way to represent uncertainty about represented values. An animal must frequently act on the basis of incomplete or ambiguous sensory data about its environment. Wiser decisions are possible if the animal's representation of the environment includes not only best estimates of environmental signals, but also information about the certainty of those estimates. One decision an animal might make is to collect more information, in order to reduce uncertainty to the point where a more critical decision can be made.

The problem is not unique to perception. Even if the instantaneous value of a signal is unambiguous, an animal sometimes has to act on its predicted future value. For example, Jane Goodall [137] relates an incident in which a young chimpanzee died while trying to jump from one high tree branch to another, because a gust of

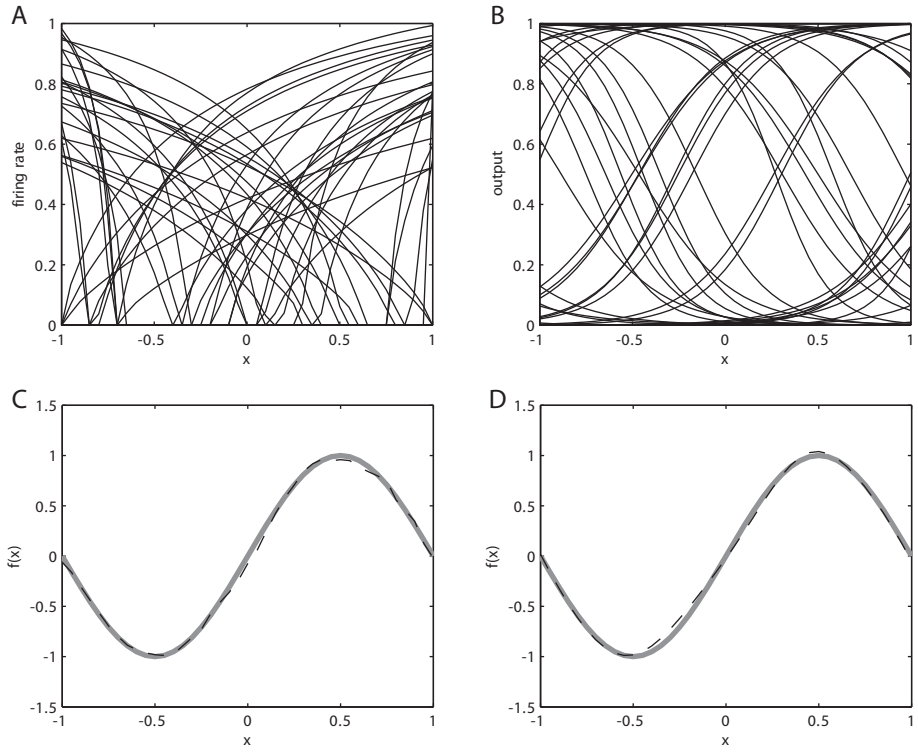

Figure 3.1: Computation with a population of neurons that are tuned to a one-dimensional signal (left panels) is analogous to computation with a feedforward artificial neural network with one input unit (right panels). A, Firing rates (normalized) of a population of neuron models, as a function of a one-dimensional encoded signal. These neuron models have randomly-selected parameters, which produce uniform distributions over ranges of threshold and peak firing rate. B, Sigmoidal responses of various hidden-layer units in a feedforward artificial neural network, as a function of a single input. These units have randomly-selected input weights and biases. C, A function (solid line) of the encoded signal, and its approximation (dashed line) as a linear combination of the tuning curves in (A). D, Function of the input (solid line) and approximation (dashed line) by an output unit, as a linear combination of the tuning curves in (B).

wind arose mid-jump. The brain of an older chimpanzee might have contained a more accurate representation of this uncertainty, and made a different decision.

In theory, a single-neuron code could provide information about both a parameter estimate and its uncertainty, in different frequency bands. However, population codes allow more sophisticated representations of uncertainty. This is illustrated by several types of stochastic artificial neural networks, in which binary nodes have a certain probability of being active in each time step (e.g. the Boltzmann machine [157]). These networks learn to model the probability distribution of their inputs, in such a way that novel inputs can be classified stochastically. Interestingly, these networks can also generate novel patterns that conform to a certain category of input.

Other network models have been devised in which *instantaneous* activity encodes a probability distribution. The first was due to Anderson [21, 22], who proposed a variation on the population vector model [130], in which each neuron's activity corresponds to a probability density function (PDF) rather than a vector. Population activity can then be interpreted as a composite PDF, which is the normalized sum of each neuron's PDF, weighted by the corresponding neuron's firing rate. These ideas are among the roots of the Neural Engineering Framework, which is discussed below. The constraint that each neuron's activity corresponds to a PDF turns out to be unnecessary, and the model has since been refined so that individual neurons contribute more general basis functions to a composite PDF at the population level [39]. Other related models have also been developed (e.g. [321, 401]).

3.2 Neural Engineering Framework

The population-coding concepts described above were integrated by Eliasmith & Anderson [111] into a coherent theoretical framework, called the Neural Engineering Framework (NEF). The term "neural engineering" has since come to be strongly associated with brain-machine interfaces. In the NEF, the term is instead used to emphasize that the brain solves practical problems using practical (i.e. noisy, saturating, failure-prone) components, and that consequently, many of the analytical methods that have been developed to solve engineering problems are well-suited for studying neural systems. The framework is based on the three following principles (their wording):

1. Neural representations are defined by the combination of nonlinear encoding and weighted linear decoding.

2. Transformations of neural representations are functions of variables that are represented by neural populations. Transformations are determined using an alternately weighted linear decoding.

3. Neural dynamics are characterized by considering neural representations as control theoretic state variables.

The central concept in this framework is that of neural representation, i.e. that the activity of a population of neurons "represents" something. Population activity may represent a fact about the physical world (such as a sound, the location of part of the body, etc.), or an abstract object that is relevant to information processing. Mathematically, the framework describes representation of scalars, vectors, and functions, which together can be used to model a great variety of physical and abstract entities.

The physiology of representation is probably most intuitive in terms of the senses. The transducer cells of sensory organs (e.g. retina, basilar membrane, muscle spindles) are connected to neurons. The firing patterns of these neurons are determined by the signals (e.g. light, sound, muscle length) to which these cells are sensitive. Furthermore, to the extent that the nervous system is sensitive to differences between signals (e.g. between two light levels), different signals produce different overall patterns of activity in the group of neurons that are connected to the sense organ. So the activity of these neurons contains encoded information about the physical world. Primary sensory neurons are connected to deeper neurons, allowing sensory information to propagate through the brain. Different patterns of connectivity can result in different transformation of sensory information, or its combination with information from other sources.

3.2.1 Representation

The NEF theory of neural representation combines prior work on 1) decoding single spike trains through linear filtering [314], and 2) optimal linear decoding of population vectors [1, 323], and generalizes the description of representation beyond scalars and vectors to include representation of functions. For simplicity, the discussion will focus on vector representation (scalars are just one-dimensional vectors, and functions are just infinite-dimensional vectors, which can be approximated arbitrarily well as high-dimensional vectors).

There are two aspects of representation: encoding and decoding. The description of encoding begins with the current J_i that enters each neuron as a function of the represented vector $\mathbf{x}$ (ignoring for the moment the physiological source of this current):

$$J_i(\mathbf{x}) = \alpha_i \tilde{\phi}_i^T \mathbf{x} + J_i^{bias},$$

where α_i is a scale factor, $\tilde{\phi}_i$ is an encoding unit vector (e.g. -1 or 1 in the scalar representation case), and J_i^{bias} is a "bias current", which models all the effects that contribute to baseline firing, in the absence of input. Neural activity a_i is a function $G(\bullet)$ of current,

$$a_i(\mathbf{x}) = G_i[J_i(\mathbf{x})] + \eta_i(t),$$

where $\eta_i(t)$ summarizes the various sources of noise that affect the neuron's output. This equation determines the neuron's tuning curve, i.e. its activity as a function of the represented variable $\mathbf{x}$. $G_i(\bullet)$ is generally a non-linear, non-decreasing function of $J_i(\mathbf{x})$. This means that for a given magnitude of $\mathbf{x}$, the neuron's activity is greatest when $\mathbf{x}$ is aligned with the neuron's encoding vector (or preferred-direction vector) $\tilde{\phi}_i$. This form of tuning is called cosine tuning, because $\tilde{\phi}_i^T \mathbf{x} \propto cos(\theta)$, where θ is the angle between $\tilde{\phi}_i$ and $\mathbf{x}$. Cosine tuning is discussed further in Section 3.3.

Depending on the level of detail of the model, $a_i(\mathbf{x})$ can itself be a scalar that corresponds to the neuron's firing rate, or it can be a more detailed spike-based description, for example a sum of impulses,

$$a_i(\mathbf{x}, t) = \sum_n \delta(t - t_{in}),$$

where t is time and t_{in} is the time at which the i^{th} neuron spikes for the n^{th} time. Models of a neuron's spike-response function $G(\bullet)$ range from simple sigmoid functions to conductance models with dozens of parameters. For models of large networks, the leaky-integrate-and-fire (LIF) model [207] provides an appealing balance between fidelity and complexity. This model is simple, but it emulates several key features of spiking behaviour (e.g. all-or-nothing spikes; spike-rate saturation). The firing rate of an LIF neuron is

$$a_i(\mathbf{x}) = \frac{1}{\tau_{ref} - \tau_{RC} \ln(1 - J_{th}/J_i(\mathbf{x}))},$$

where τ_{ref} is the minimal refractory time between spikes, τ_{RC} is the membrane time constant, and J_{th} is the current threshold at which the firing rate becomes non-zero. With minor elaborations, the LIF model can predict the spike times of individual recorded neurons very accurately (sometimes outperforming much more complex conductance models [275]).

Completing the account of representation, decoding is described in terms of a weighted sum of the activities of neurons in the population,

$$\hat{\mathbf{x}} = \sum_i a_i(\mathbf{x}) \phi_i,$$

where ϕ_i are decoding vectors, and $\hat{\mathbf{x}}$ is the decoded estimate of the input $\mathbf{x}$. Least-squares optimal decoding vectors can be found by minimizing the following error:

$$E(\Phi) = \frac{1}{2} \int_{\mathbf{x}} [\mathbf{x} - \sum_i^N a_i(\mathbf{x}) \phi_i]^2 d\mathbf{x}.$$

As discussed in the introduction, one advantage of population coding is the possibility of explicitly representing the uncertainty of an estimate. Representation of uncertainty, in the form of a probability distribution, is a special case of function representation [39].

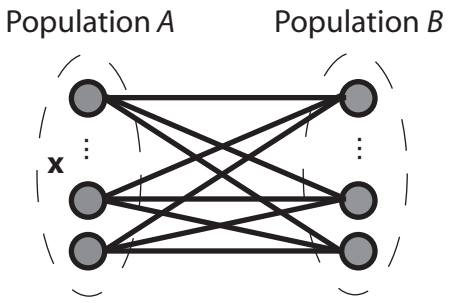

Figure 3.2: A communication channel between a population of neurons A and another population B.

3.2.2 Transformation

If a population A represents the signal $\mathbf{x}$, then it can convey the value of $\mathbf{x}$ to another population B (Figure 3.2), provided the synaptic weights between populations are tuned appropriately. To find the synaptic weights that support this communication, we begin by assuming that population B does not have direct access to the value of $\mathbf{x}$, but that it has access to $\hat{\mathbf{x}}$, the estimate of $\mathbf{x}$ that can be decoded from the activity of neurons in A. The activity b_j of the j^{th} neuron in population B is then

$$\begin{aligned} b_j(x) =& G_j[J_i(\hat{\mathbf{x}})] \\ =& G_j[\alpha_j \tilde{\phi}_j^T \hat{\mathbf{x}} + J_j^{bias}] \\ =& G_j[\alpha_j \tilde{\phi}_j^T \sum_i a_i(\mathbf{x})\phi_i + J_j^{bias}] \\ =& G_j[\sum_i w_{ji} a_i(\mathbf{x}) + J_j^{bias}] \end{aligned}$$

where $w_{ij} = \alpha_j \tilde{\phi}_j^T \phi_i$ is the weight of the i^{th} presynaptic neuron's synapse onto the j^{th} post-synaptic neuron. In other words, the weight of the synapse from the i^{th} presynaptic neuron onto the j^{th} postsynaptic neuron is the product of the corresponding preferred-direction and decoding vectors, scaled by α_j (which is a scale factor common to all the synapses onto the j^{th} neuron). Note that the term $\sum w_{ji} a_i(\mathbf{x})$ describes linear synaptic integration, i.e. the inputs to a neuron combine linearly before passing through the output nonlinearity G. Thus the NEF characterization of linear decoding and cosine tuning implies linear synaptic integration. As discussed in Chapter 8, the NEF characterization of cosine tuning is also implied *by* linear synaptic integration.

With a different choice of synaptic weights, population A can communicate a function $f(\mathbf{x})$ to population B, rather than $\mathbf{x}$ itself. For example, if all the weights are doubled, the projection communicates $f(\mathbf{x}) = 2\mathbf{x}$. More generally, a linear

transformation defined by the matrix A is communicated by the synaptic weights

$$w_{ij} = \alpha_j \tilde{\phi}_j^T A \phi_i.$$

Non-linear transformations can also be performed, *via* alternate decoding vectors $\phi_i^{f(\mathbf{x})}$, which optimally decode $f(\mathbf{x})$ instead of $\mathbf{x}$.

3.2.3 Dynamics

For neurons that fire at a constant rate with constant injected current, firing rate dynamics are dominated by the dynamics of post-synaptic current [111]. A spike at a chemical synapse results in post-synaptic current that has a time course much like a first-order exponential decay,

$$h(t > 0) = \frac{w}{\tau} e^{-t/\tau},$$

where w is a synaptic weight (which describes the total charge entering the cell with each spike), and the time constant τ describes the rate of current decay. This model of post-synaptic current (PSC) dynamics ignores a small but finite rise time, but this simplification does not substantially affect the present discussion. The net current at each synapse is the convolution of presynaptic spike impulses with $h(t)$. Assuming linear synaptic integration, this implies that the total input to the cell is the weighted sum of the presynaptic population output, all convolved with $h(t)$ (Figure 3.3).

Represented variables can be treated as the state variables of a dynamical system. A linear time-invariant system with a finite list of states $\mathbf{x}$ is described by the following equations:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= A\mathbf{x}(t) + B\mathbf{u}(t), \\ \mathbf{y}(t) &= C\mathbf{x}(t) + D\mathbf{u}(t), \end{aligned} \tag{3.1}$$

where $\mathbf{u}(t)$ and $\mathbf{y}(t)$ are the system input and output, respectively, and A, B, C, and D define the system's dynamics, and its input, output, and feed-through scaling, respectively. These equations can be re-written in terms of PSC dynamics, so that the activity of a population with an input $\mathbf{u}(t)$ and a recurrent feedback connection can be described in a form similar to the first state equation:

$$\mathbf{x}(t) = \frac{1}{\tau} e^{-t/\tau} * [A'\mathbf{x}(t) + B'\mathbf{u}(t)], \tag{3.2}$$

where A' and B' describe the recurrent and input projections, respectively. If equations (3.1) and (3.2) are re-written in the Laplace domain, as follows:

$$s\mathbf{x}(s) = A\mathbf{x}(s) + B\mathbf{u}(s),$$

$$\mathbf{x}(s) = \frac{1}{\tau s + 1}[A'\mathbf{x}(s) + B'\mathbf{u}(s)],$$

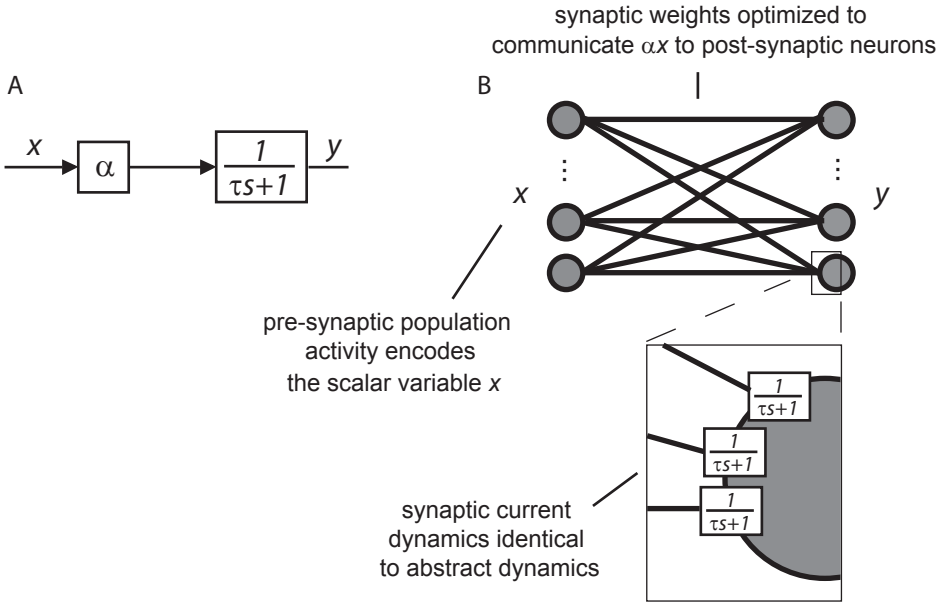

Figure 3.3: Transformations and dynamics of represented variables vs. neuron activities. A, Block diagram describing the relationship between two hypothetical represented variables, x and y. The diagram includes a static block, which multiplies x by a constant α, and a dynamic block, which filters αx to produce the output y. B, Schematic of a neural network model that corresponds to this block diagram. Each circle indicates a neuron. On the left is a population of neurons that encodes the value of x (i.e. the firing of these neurons is correlated with x, in such a way that x can be estimated from their activity pattern). These neurons project to another population of neurons on the right, which encodes the value of y. The synaptic weights in this projection are chosen so that a decoded estimate of αx is conveyed to the y neurons (as described in Section 3.2.2). The transfer function of each synapse, which maps a spike input to a post-synaptic current response, is the same as the transfer function of the block diagram in (A). As a result of these similarities, the neural network behaves analogously to the abstract system shown in (A).

then they are the same when

$$A' = \tau A + I,$$

$$B' = \tau B.$$

Therefore, 1) every linear time-invariant system has an equivalent family of neural circuits (to within coding error), which differ in terms of time constants, firing rates, etc., and 2) it is straightforward to find this family. Although the above discussion is restricted to linear dynamics, the same mechanisms can also implement nonlinear systems, through nonlinear function decoders $\phi_i^{f(x)}$ [110]. In principal, the result applies to any dynamical system that can be described with a set of explicit ordinary differential equations, although (as discussed in Section 3.3, below) not all non-linear functions can be decoded with equal accuracy.

3.2.4 Summary

The NEF provides a broad and coherent account of how neurons work together in large circuits to represent and transform information. When a circuit model is developed with the NEF, the modeler specifies low-level neuronal details such as ranges of firing rates and membrane time constants, as well as high-level details relevant to the population code, such as the preferred direction vector of each neuron. Probability distributions over these parameters can be estimated from results in the experimental literature. The modeler also specifies the high-level transformations performed by each projection from one population to another. These transformations are hypothesized based on experimental evidence about the computational and dynamic properties of the circuit, and its function.

Finally, the synaptic weights necessary to produce the specified transformations are derived as a function of everything else. This is a great advantage, because synaptic weights are much more difficult to measure than other circuit properties. Furthermore, even if a few synaptic weights can be measured, providing a rough estimate of their probability distribution, such a distribution indicates little about the key property of the synaptic weights, which is their fine structure in relation to neuronal tuning curves.

In contrast with the NEF, most approaches to neural modeling require synaptic weights to be learned. This is more time-consuming, and it can be difficult to find a learning rule that leads to the apparent function of a neural circuit. On the other hand, synaptic weights are learned in the brain, so the requirement that they be learned in a model can arguably provide an additional constraint. However, theoretical learning rules are much less sophisticated than physiological mechanisms of synaptic plasticity (which are a subject of very active research). So it is unclear how well the constraints imposed by theoretical learning rules are aligned with the constraints of biology. Recognizing that this alignment is likely to improve in the future, Chapter 8 shows how learned weights can be interpreted within the NEF.

In a population code, the codebook consists of neuronal tuning curves $a_i(\mathbf{x})$, i.e. firing rates as functions of represented variables. The shapes and distribution of tuning curves define both the redudancy inherent in the code, and the computations that it supports. Furthermore, as the NEF shows, these computations also underlie network dynamics. So tuning curves are key features of a neural circuit. At first glance, cosine tuning seems to imply that tuning curves belong to a fairly restricted family, and one might suspect that this entails some sort of restriction on computation. The following section explores this issue in more detail.

3.3 Cosine Tuning

Linear decoding, which corresponds to linear synaptic integration, constrains transformations to be weighted sums of the tuning curves of the presynaptic population. The shapes of the tuning curves therefore determine which transformations are possible.

The tuning curves of a population are typically diverse. They are therefore something like a vector-space basis, in that different linear combinations of tuning curves produce a variety of functions. However, the components of a basis are linearly independent, by definition, whereas neuronal tuning curves are typically not (it is this redundancy that mitigates the effects of noise and damage). So the tuning curves of a neuronal population do not form a basis, but a frame [90].

A frame is a list of vectors $\Psi = \{\psi_k\}$ in a vector space H, for which

$$\alpha||\mathbf{v}||^2 \leq \sum_k |<\mathbf{v}, \psi_k>|^2 \leq \beta||\mathbf{v}||^2 \tag{3.3}$$

for all $\mathbf{v}$ in H, where $<\bullet,\bullet>$ denotes an inner product, and $\alpha > 0$ and $\beta < \infty$ are called frame bounds. The lower bound means that Ψ spans H, and the upper bound means that Ψ is finite.

A vector $\mathbf{v}$ (e.g. a sampled function of an encoded variable x) can be encoded on a frame Ψ by the frame operator F. The frame operator is defined by

$$c_k = (F\mathbf{v})_k =<\mathbf{v}, \psi_k> .$$

The representation of $\mathbf{v}$ on the frame is then $\mathbf{c} = F\mathbf{v}$. In the context of a neuronal population code, ψ_k corresponds to the k^{th} tuning curve, and c_k corresponds to the inner product of the k^{th} tuning curve and $\mathbf{v}$. This inner product has no clear physiological meaning, so encoding on a frame is quite different from neural encoding (although as discussed in Section 3.3.2, c_k takes on meaning in the context of Hebbian learning). The vector $\mathbf{v}$ can be decoded from its frame representation $\mathbf{c}$ as

$$\mathbf{v} = \tilde{F}^T\mathbf{c} = \sum_k c_k\tilde{\psi}_k,$$

where $\tilde{\psi}_k$ make up the "dual" frame of Ψ, and $\tilde{F}$ is the corresponding frame operator. The dual is defined by

$$\tilde{\psi}_k = (F^T F)^{-1} \psi_k.$$

Equation (3.3) can also be written as

$$\alpha I \leq F^T F \leq \beta I. \tag{3.4}$$

The space of all possible functions of a represented variable is infinite-dimensional, so of course no finite list of tuning curves can span it. Consequently, the transformations that can be realized by a projection from any given population are constrained, i.e. to the finite-dimensional space that is spanned by its tuning curves. The principal components of the tuning curves form a basis of this space. However, the dimension of the space is not necessarily clear-cut. It is not uncommon for a few principal components to account for most of the variance of the tuning curves, and for many additional principal components account for the remaining variance. Ideally, the space of possible transformations would include all of the principal components that contribute to the tuning curve variance. However, neurons are noisy, and principal components with variance comparable to or smaller than the noise cannot be accurately decoded. So roughly speaking, the dimension of the space is equal to the number of principal components that account for more of the variance in neuronal activity than noise does.

As mentioned above, linear synaptic integration implies cosine tuning. Therefore the principal components of cosine-tuned neurons are of particular interest. Interestingly, the principal components of cosine-tuned LIF neurons closely resemble the Legendre polynomials [111] (see also Figure 3.4). This has led C. H. Anderson (personal communication) to argue that it should be possible to understand neural transformations largely in terms of low-order polynomial functions of represented variables.

3.3.1 Cosine Tuning on a Manifold

While there are clear examples of cosine tuning in the brain (e.g. [130]), Gaussian tuning is also frequently described. ("Gaussian" tuning curves are not literally Gaussian functions, but they are similar in that they are smooth, localized, and symmetric.) Since linear synaptic integration implies cosine tuning, Gaussian tuning appears at first to imply non-linear synaptic integration. However, one-dimensional Gaussian tuning also describes two-dimensional cosine-tuned neurons, if the experiment only includes stimuli that fall on a circle in the two-dimensional space (as in e.g. [246]).

Similarly, Gaussian tuning can arise from cosine tuning on a manifold that arises from transformations in the neural circuit. Such a manifold arises when an n-dimensional population receives m-dimensional information, where $m < n$. Figure 3.5B shows an example of two-dimensional cosine tuning that appears Gaussian on

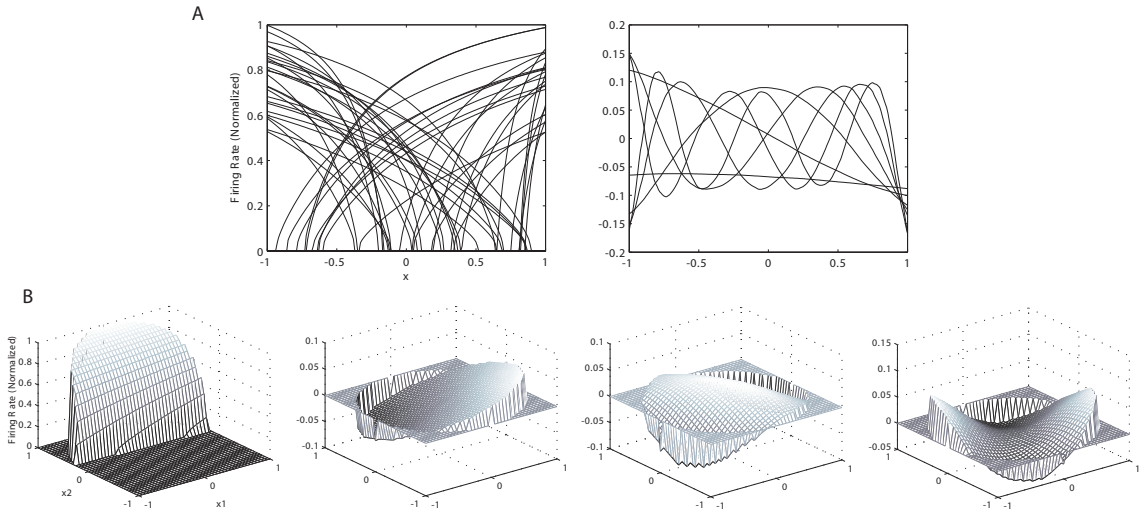

Figure 3.4: Principal components of cosine-tuned LIF neurons resemble low-order polynomials. A, Representative tuning curves from a population of LIF neurons that are tuned to one dimension (left), and the first seven principal components of the tuning curves of this population. B, A single example of the tuning curve of an LIF neuron that is tuned to two dimensions (left), and three examples of principal components of a population of such neurons (right).

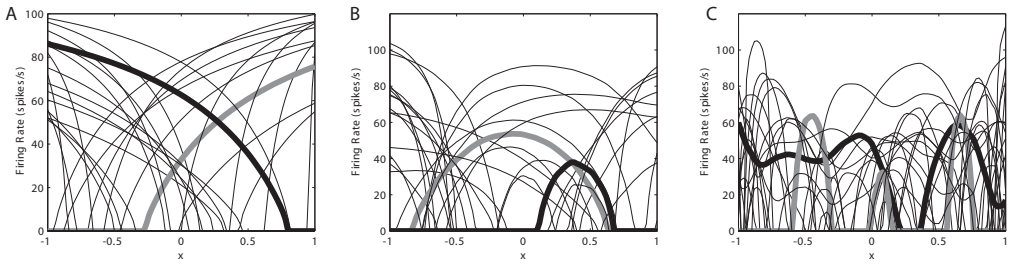

Figure 3.5: Cosine tuning on manifolds. Two example tuning curves are shown in bold in each panel. A, Example tuning curves of one-dimensional LIF neurons. B, Tuning curves of two-dimensional cosine-tuned LIF neurons on a one-dimensional manifold that is made up of the second and third principal components of the tuning curves in (A). Many of these resemble Gaussian functions. C, Tuning curves of nine-dimensional cosine-tuned LIF neurons on a one-dimensional manifold that is made up of principal components 2-10. These tuning curves are irregular and often multi-modal.

a one-dimensional manifold. In this example, a one-dimensional population drives a two-dimensional population with transform

$$f(x) = \begin{bmatrix} p_2(x) \\ p_3(x) \end{bmatrix},$$

where x is a scalar variable represented by the presynaptic population, and p_i are the second and third principal components of the presynaptic population's tuning curves (the first is omitted because it is essentially constant). The neurons in the post-synaptic population are driven by one-dimensional information, so their activity is restricted to a one-dimensional manifold. The tuning curves of this population can therefore be plotted in the one-dimensional space, as functions of x, instead of along the one-dimensional manifold in the two-dimensional space. In this space they appear roughly Gaussian.

Manifolds in higher-dimensional spaces can also be considered. Figure 3.5C illustrates a projection from a one-dimensional population onto a manifold that follows the first ten principal components of its tuning curves. In this high-dimensional space, if each neuron has a different preferred direction, then it decodes a unique function of the represented variable. When the preferred directions are drawn randomly from the unit hypersphere, the resulting tuning curves are not generally Gaussian, but have irregular shapes. Chapter 5 explores irregular tuning curves like these in more detail, and illustrates how they can evade experimental detection while seriously confounding the interpretation of electrophysiological data. Although these tuning curves seem exotic, they are nothing more than the result of cosine tuning on the principal components of a one-dimensional population.

3.3.2 Tight Frames

If $\alpha = \beta$ in (3.3), then the frame is "tight". Intuitively, a tight frame spans the vector space evenly, i.e. it does not span any part of the space more densely than other parts. An orthogonal basis is the smallest kind of tight frame; an orthonormal basis has frame bounds $\alpha = \beta = 1$.

Tight frames are of special interest in neuroscience, because a projection from a population that forms a tight frame can be trained to calculate any function of the represented variable, *via* localized Hebbian learning [92]. In Hebbian learning, synaptic weights are modified in proportion with the product of presynaptic and post-synaptic activity. In Hebbian learning of a function on a tight frame, the post-synaptic neurons must be driven in the correct patterns during learning, but no error or reward signal is needed.

This works because for a tight frame, from (3.4), $F^T F = \alpha I = \beta I$, so that

$$\tilde{\psi}_k = (\beta I)^{-1} \psi_k = \frac{1}{\beta} \psi_k.$$

So a function $\mathbf{v}(x)$ can be produced from ψ_k as

$$\mathbf{v} = \sum_k c_k \tilde{\psi}_k = \frac{1}{\beta} \sum_k < \mathbf{v}, \psi_k > \psi_k.$$

These coefficients scale with $< \mathbf{v}, \psi_k >$, independently of $\psi_{i \neq k}$. If ψ_k are presynaptic tuning curves, and $\mathbf{v}(x)$ is post-synaptic activity, then this inner product over the coding space is approximately equal to the integral of the instantaneous product of presynaptic and post-synaptic activity, as different x are sampled over time. This is the product that results from Hebbian learning.

Cosine-tuned LIF neurons do not form a tight frame. In fact, there are few large principal components, and the variance associated with more minor components decreases smoothly, so that the lower frame bound is very small. Gaussian tuning curves have more and different principal components than cosine tuning curves [379]. However, although they form higher-dimensional frames, these frames are also not tight. The same is true for cosine-tuned neurons on higher-dimensional manifolds, when the encoding vectors are evenly distributed. However, interestingly, it is possible to find sets of encoding vectors that form a relatively tight frame on a manifold that is made up of several principal components of one-dimensional LIF tuning curves (Figure 3.6).

3.4 Discussion

The Neural Engineering Framework provides a coherent view of the representation of scalars, vectors, and functions, as either ratios between different firing rates, or

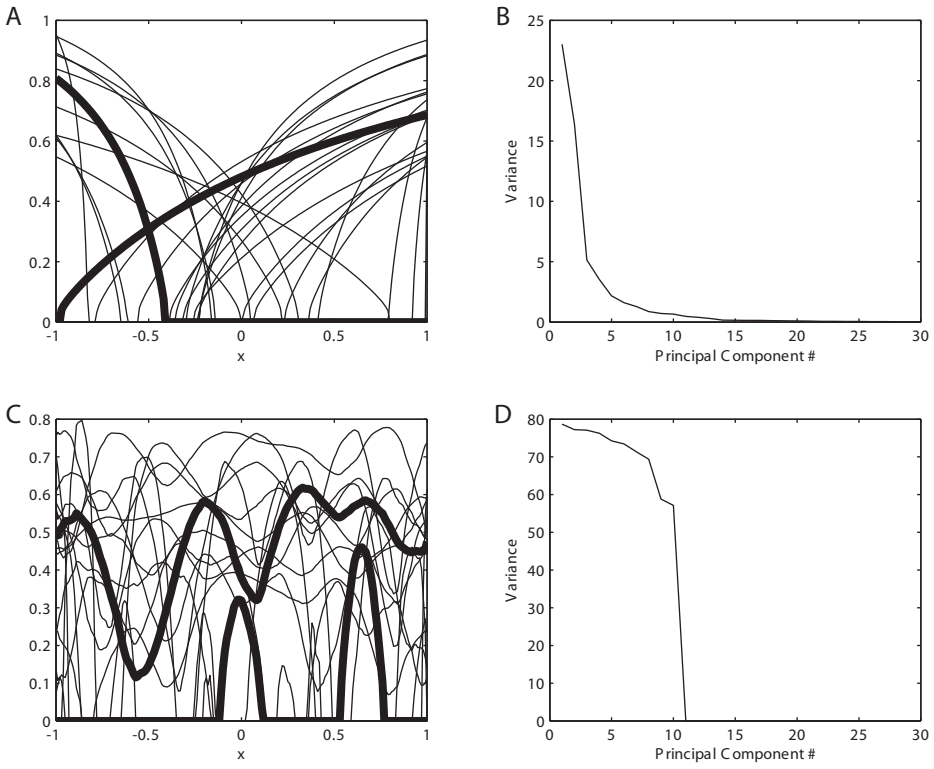

Figure 3.6: Tightness of LIF neurons on a one-dimensional space (A&B) and on a one-dimensional manifold in a ten-dimensional space (C&D). The left panels (A&C) show 30 representative tuning curves drawn from 300-neuron populations, with selected examples in bold. The right panels (B&D) show the variance of the population tuning curves that is accounted for by each principal component. A small number of dimensions account for most of the variance of the tuning curves in the one-dimensional space (top), but there is no clear answer to the question of how many dimensions are spanned by these neurons. Some are spanned much more densely than others. In contrast, the tuning curves on the manifold clearly span a ten-dimensional space, and they span this space relatively evenly.

temporal patterns of action potentials. It also generalizes this view to explain how information is communicated and transformed in projections between populations of neurons. Finally, it explains how such transformations combine with synaptic properties to determine large-scale network dynamics. The framework clearly defines the relationships between different observable properties of a neural system, e.g. response functions of individual neurons, and high-level network behaviour. This allows the modeler to integrate diverse sources of experimental data.

Despite its considerable scope, the framework does not approach the sophistication of biological neural systems. For example, fundamental to the operation of any neural system are the mechanisms of synaptic plasticity that determine synaptic strengths. From one perspective, the NEF avoids a great deal of complexity by deriving synaptic weights that give rise to observed or hypothesized network behavior. From another perspective, plasticity is itself an important aspect of network behaviour that the NEF ignores. Eliasmith & Anderson [111] touch on the question of plasticity by showing how a communication channel could be learned, but this brief treatment leaves many questions unanswered.

There are also a number of subtleties of neuron behaviour that the NEF ignores, for the sake of simplicity. For example, while the NEF assumes linear synaptic integration, there are many clear examples of nonlinear synaptic integration in the brain, and there are some reasons to think that nonlinear synaptic integration may be quite common. Spiking dynamics are another example. The NEF assumes that neuron dynamics are dominated by the dynamics of post-synaptic current. This is true for some neurons, e.g. fast-spiking interneurons, but it is not true for many others. In many neurons, the spiking process itself has prominent dynamic properties, such as (in various different neuron types) firing-rate adaptation, intrinsic bursting, and hysteresis.

One useful property of the NEF is that it unifies spike-rate codes and spike-timing codes. A great deal of attention has been focused in the literature on whether, in various systems, the precise timing of spikes (on the scale of a few milliseconds) contains additional information beyond that contained in firing rates (i.e. the number of spikes over longer time windows, e.g. 100 ms). In the NEF, the neural code is a continuum between rate and timing codes. When firing rates are high, and represented variables change slowly, the neural code resembles a rate code (i.e. most of the information is in the firing rates). On the other hand, when the represented variable changes quickly and spike rates are low, the neural code resembles a timing code (i.e. most of the information is contained in spike timing). In this view, a timing code is essentially the rate code of a precisely-timed signal. However, the brain may contain more subtle forms of timing code, including timing codes for slowly-changing signals.

All of the above caveats are relevant to the basal ganglia. Basal ganglia neurons have non-linear dendritic integration properties, prominent non-linear spike dynamics, plastic synapses, and irregular spike timing patterns that correlate with behaviour. In a sense, most of the remaining chapters constitute a generalization of

the NEF in these directions, i.e. an account of how timing codes, nonlinear spiking dynamics, and synaptic plasticity interact with the basic NEF theory. The set of basic population-coding principles provided by the NEF, extended to account for further details of basal ganglia physiology, should be very useful for understanding basal ganglia function.

Furthermore, these properties are not unique to the basal ganglia, but are also prominent in the cortex, and in other subcortical areas. So although the present goal is to elaborate the NEF so that it can account more thoroughly for basal ganglia physiology, many of the results apply very broadly to the diverse neural systems that employ population codes.

Chapter 4

Non-Linear Synaptic Integration

The NEF assumes linear decoding. This corresponds to the conservative assumption of linear synaptic integration, which is shared by the majority of network models. However, synaptic integration in real neurons involves a number of strongly nonlinear processes, including saturation effects, dendritic spikes, and non-linear inter-branch interactions. This chapter discusses three types of nonlinearity that can have a large impact on synaptic integration: 1) nonlinear inter-branch interactions, 2) shunting conductances, and 3) supralinear input-conductance relationships within a dendritic branch. All of these nonlinearities are found to have an impact on neural representation. The first type of nonlinearity influences the encoding process, by shaping the neuron's response (tuning curve) to linearly decoded information. In contrast, the latter two nonlinearities support alternative decoding mechanisms. Like linear decoding, these nonlinear mechanisms enable decoding of both linear and nonlinear functions of represented variables. In addition, the nonlinear mechanisms have certain advantages. It is shown that shunting inhibition allows Hebbian learning of a transform, even when presynaptic tuning curves do not form a tight frame. Intra-branch nonlinearities do not require diverse presynaptic tuning curves in order to decode a function of the input. Therefore, while challenging the NEF, the presence of such nonlinearities in real neurons also suggests ways in which it can be usefully extended.

4.1 Introduction

The NEF emphasizes linear decoding. There are two reasons for doing this: 1) it simplifies the theory dramatically, and 2) linear decoding can be performed by neurons, through linear synaptic integration. By restricting consideration to decoding methods that a neuron can perform, rather than something more powerful (e.g. Bayesian decoding [402]) the theory remains focused on the realistic information-processing capacity of neural circuits, rather than the information that an ideal

observer could extract from them.[1]

However, it is fair to question whether synaptic integration can reasonably be approximated as linear. The thin branching structure of the dendritic tree, which is unique to neurons across all cell types, increases the surface area available for synaptic contacts. However, it also effectively isolates some portions of the cell electronically from others, which raises the possibility that different inputs are integrated in more complex ways than summation. Furthermore, dendrites are not passive conduits of synaptic inputs, but contain a variety of voltage-dependent ion channels (reviewed by [235]). Interestingly, some of these channels appear to compensate for passive cable properties, rather than mediating nonlinear synaptic integration [236, 318]. For example, one common type of active dendritic current can help to compensate for low-pass filtering due to cable properties [235, 385].

In other cases, nonlinear dendritic processes result in nonlinear synaptic integration. This chapter studies three key types of nonlinear synaptic integration, and shows how they interact with computations that are based on population codes.

4.2 Inter-Branch Non-Linearity

Inputs to different branches of a dendritic tree combine nonlinearly in some neurons. A striking example is found in the medial superior olivary nucleus (MSO) of the brainstem. The majority of neurons in this nucleus have two major dendritic branches: one receives input from the ipsilateral cochlear nucleus, and the other from the contralateral cochlear nucleus [342]. The activity of neurons in the cochlear nuclei are oscillatory, and tightly phase-locked with rarefaction at the corresponding ear. The neurons of the medial superior olive fire only when inputs from each side are received at the same time [398]. This coincidence-detection mechanism, in conjunction with a network of delay lines, allows the animal to localize low-frequency sounds [184, 66]. Synaptic integration in a coincidence-detecting MSO neuron can be modeled as a product of the inputs to the two major branches. The firing rate $b(\mathbf{x})$ of the model is then

$$b(\mathbf{x}) = G[d_i d_c],$$

where the function $G(\bullet)$ maps net synaptic current onto the firing rate, and d_i and d_c, the net inputs to the ipsilateral and contralateral branches, respectively, are

$$d_j = \sum_k w_{jk} a_{jk}(x_j(t)) * \delta(t - \Delta_j).$$

Here x_j are the rarefaction signals represented by the cochlear nuclei, $a_{jk}(x_j)$ are the activities of the neurons in these nuclei, w are synaptic weights, δ is the delta

[1]Note that this is separate from the issue of whether neural circuits can be designed which decode inputs in a more sophisticated manner. Usually, the purpose of neural circuits is not to decode inputs. Instead, decoding is an abstraction that describes part of the process of neural computation.

(impulse) function, and Δ_j is the conductance delay of the projection from the j^{th} cochlear nucleus to the corresponding dendritic branch.

This model can be interpreted in two ways. One interpretation is that the neuron decodes a non-linear function of its two-dimensional input $\mathbf{x}$. From this perspective the nonlinearity contributes to computation. However, an alternate interpretation is that the neuron decodes $\mathbf{x}$, and then encodes this information in a non-linear tuning curve. From this perspective, the nonlinearity contributes to representation.

The diversity of responses across a population determines which perspective is more useful. The responses of different MSO neurons have essentially the same nonlinearity. This means that the product of the two inputs (i.e. $x_1 x_2$) is well-represented by a population of these neurons, but that the representation of the input $\mathbf{x}$ is poor in general. So in this case, the first interpretation is more reasonable. However, if the responses of different neurons were more diverse, then it would be difficult to associate them with a single computation, but they would form a rich code of their input. In that case the latter interpretation would be more reasonable.

An example of the latter case can be found in the medium spiny neurons of the striatum. These neurons integrate inputs in a more subtle manner, such that the neuron does not fire unless the membrane potential is elevated in several of the roughly two dozen main branches [388]. This nonlinearity can be illustrated with a simple phenomenological model. In this model, each medium spiny neuron (MSN) has four main branches, two of which must have an elevated membrane potential in order for the neuron to fire. Synaptic integration is linear within each branch, and each of the four branches is tuned to a different preferred direction in a two-dimensional space (a two-dimensional space is used in order to allow visualization of the tuning curves). The activity of the i^{th} neuron is described by the tuning function

$$a_i(\mathbf{x}) = G[m_b(\tilde{\phi}_b^T \mathbf{x})],$$

where the response function G is sigmoidal, $\tilde{\phi}_b$ is the encoding vector of branch b, and $m_b(\bullet)$ is the minimum over the two most active branches (a fuzzy "and"). The sigmoid function has a steep transition, modelling the tendency of these neurons to self-stabilize in "up" and "down" states (although the well-known clear separation of these states *in vivo* may have been an artefact of synchrony in cortical input, due to surgical anaesthesia [239]). Figure 4.1 shows examples of the resulting tuning curves. These tuning curves are diverse, and form a rich representation of the input. The only difference between this representation and a representation arising from linear synaptic integration is the shapes of the tuning curves.

The above examples involve relatively specialized neurons with prominent non-linearities. However, cortical pyramidal cells (the most numerous cells in the cortex) also appear to combine inputs from different branches nonlinearly [254]. All such nonlinearities can be modeled with a nonlinear function that operates on multiple independent linear decodings, and can be understood in terms of their effects on the neurons' tuning curves.

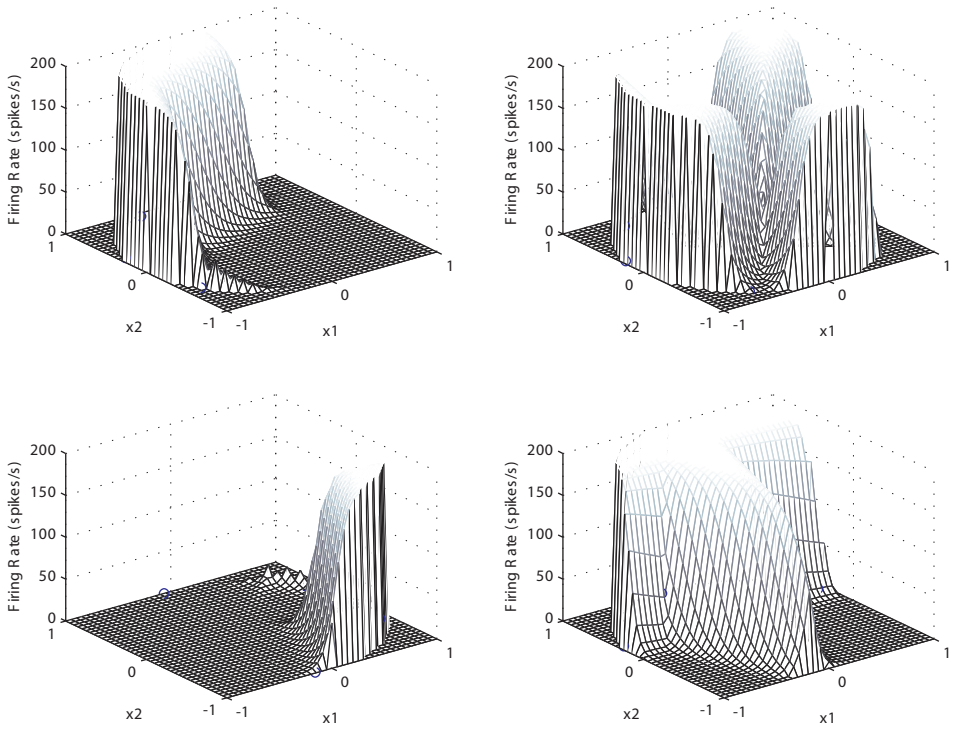

Figure 4.1: Tuning curves of selected phenomenological models of medium spiny neurons. Values outside the represented area (the unit circle in this case) are plotted as zero. The shapes depend on the distribution of preferred directions of different branches, and on how strongly the input drives each branch. The tuning curve on the bottom left is roughly sigmoidal, because the preferred directions of multiple branches are closely aligned. The tuning curves of this model can also be multimodal (top right), and can have concave (top left) or convex (bottom right) firing thresholds.

In contrast, the nonlinearities discussed in the remaining sections affect decoding, and have more profound implications.

4.3 Conductance-Current Non-Linearity

The trans-membrane current that results from the opening of a synaptic ion channel is not a linear function of channel conductance, but the product of that conductance with the difference between the membrane potential and the ion's reversal potential. For excitatory Na^+ channels, the reversal potential (about 50mV) is far from the range between rest and spike threshold (typically -65mV to -50mV), so the conductance-current relationship is only mildly sub-linear in this range. This is not the case for the Cl^- channels associated with inhibitory GABA synapses. These have a reversal potential of about -70mV, which is close to the resting potential of most cells. For this reason, the main effect of inhibition is to shunt excitatory currents, by increasing membrane conductance.

Theoretical interest in shunting inhibition derives from its divisive effect on the cell's membrane potential. This effect can be seen in the subthreshold regime of a leaky-integrate-and-fire (LIF) model, with inhibitory synaptic conductance g_i and excitatory electrode-injected current I_e. The membrane potential V evolves as

$$\dot{V} = \frac{1}{C_m}[I_e - g_i(V - E_i) - g_L(V - E_L)],$$

where g_L is the leak conductance, E_L is the reversal potential of the leak conductance, and $E_i \approx E_L$ is the reversal potential of the inhibitory channels. At equilibrium, the membrane potential is

$$V = E_L + \frac{I_e}{g_i + g_L}.$$

Thus for larger inhibitory conductances ($g_i > g_L$), the excursion of the membrane potential above rest is approximately divided by the inhibitory conductance.

All else being equal, dividing the equilibrium membrane potential corresponds roughly to dividing the cell's firing rate, so it was initially thought that shunting inhibition might provide a substrate by which neural circuits could divide one represented value by another. In a well-known example, Carandini & Heeger [65] used a shunting model to account for several properties of V1 cells. In this model, the firing rate was modelled abstractly as a function of equilibrium potential.

Unfortunately for these models, all else is not equal. In addition to dividing the equilibrium potential, shunting inhibition also divides the membrane time constant. Thus excitatory input in the presence of shunting causes a faster climb to a lower equilibrium. Using the LIF model, various authors [159, 92] have shown that one effect cancels the other, so that the net effect of shunting on the firing rate is subtractive rather than divisive.

However, studies with more sophisticated compartmental models report mixed effects (e.g. [63]). Division dominates subtraction in certain regimes that involve synaptic noise and saturation of distal dendrites [104, 305]. Similarly, Chance et al. [68] discovered a role of input noise in division, by applying noisy driving current to neurons in slice preparations. In this experiment, the strength of simulated excitatory and inhibitory drive were varied together in a balanced manner. Increases in this balanced background input had little influence on mean driving current, but increased both the variability of this current, and the membrane conductance. Larger membrane-potential fluctuations increased firing rates at the low end of the range, due to noisy fluctuations above spike threshold. These fluctuations had less effect under higher excitatory drive, when the potential rose to threshold more quickly. In combination with the subtractive effect of increased conductance, this graded amplification resulted in reduced firing rate gain, with little or no net change in firing threshold.

The Appendix of this chapter illustrates that spatial parameters (which do not appear in the LIF point-neuron model) also influence the computational effect of shunting, so that certain spatial distributions of excitation and inhibition lead to division of the firing rate, even in the absence of noise.

Taken together, these results indicate that inhibitory synaptic input can either subtract from or divide a neuron's firing rate, depending on subtleties of noise, saturation, and spatial distribution of inputs. These two operations might even co-exist on the same neuron, so that one source of inhibition divides while another subtracts.

What are the functional roles of divisive shunting in neural circuits? For one, division of excitatory drive enables divisive network computations, such as gain control. Gain control plays a number of important roles in the brain (reviewed by [325]). For example, selective attention modulates the gain of orientation-tuned neurons in the visual cortex [247]. In this context, it should also be noted that divisive network computations can be performed without shunting, e.g. *via* linear decoding of a function of two variables, $f(\mathbf{x}) = x_1/x_2$.

The roles of divisive and subtractive inhibition are of particular interest for understanding computation in the basal ganglia, where most of the neurons are inhibitory. In addition to gain control computations, the next section will show that divisive shunting can also play a more general role in population coding.

4.3.1 Average-Based Decoding

This section postulates a more general role for divisive synaptic integration, in which division leads to a different way of decoding and transforming population activity, through averages rather than sums. As shown below, this average-based decoding performs similarly to linear decoding in general, but it would have certain advantages over linear decoding in some circumstances. (The present level of analysis does not reveal any obvious disadvantages.)

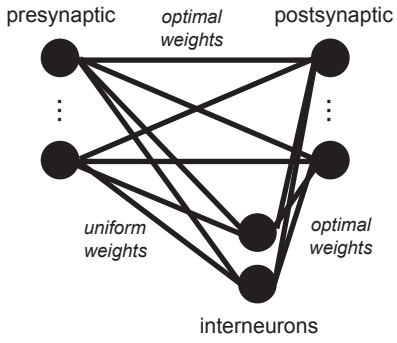

Figure 4.2: A network that communicates represented values by an averaging code. A presynaptic population (left) drives a post-synaptic population both directly, and also through a population of inhibitory interneurons, which perform divisive shunting of the direct drive. Synaptic weights in the direct projection are tuned as discussed in the text. Synapses onto the interneurons are weighted uniformly, and synapses from interneurons onto post-synaptic neurons are weighted as a communication channel (see Chapter 3).

Figure 4.2 illustrates a network that communicates through average-based decoding. In this network the presynaptic population projects to the post-synaptic population both directly, and also through a group of interneurons. This general structure is very common in the cortex, as well as the thalamus and the cortico-striatal projection, as discussed in more detail in Chapter 6. The synaptic weights in the direct projection are varied, but those in the indirect projection are uniform. The post-synaptic neurons in this network decode a function $f(\mathbf{x})$ of their input not as $\hat{f}(\mathbf{x}) = \sum_i \phi_i^f a_i(\mathbf{x})$ (as in linear decoding), but as

$$\hat{f}(\mathbf{x}) = \frac{\sum_i \bar{\phi}_i^f a_i(\mathbf{x})}{\sum_i a_i(\mathbf{x})},$$

where $\bar{\phi}_i^f$ are decoding weights for $f(\mathbf{x})$ and $a_i(\mathbf{x})$ are presynaptic activities.

The mean-squared error of this approximation of $f(\mathbf{x})$ is

$$E = \frac{1}{A}\int_{\mathbf{x}} \left[f(\mathbf{x}) - \frac{\sum_i \bar{\phi}_i^f a_i(\mathbf{x})}{\sum_i a_i(\mathbf{x})} \right]^2 d\mathbf{x},$$

where A is the size of the coded domain of $\mathbf{x}$. This error is convex, so the global minimum with respect to $\bar{\phi}_i^f$ occurs when $\partial E / \partial \bar{\phi}_i^f = 0$ for all $\bar{\phi}_i^f$. Similarly to linear decoding error (discussed by [111]), this allows us to find the vector $\bar{\Phi}^f$ of optimal $\bar{\phi}_i^f$, as

$$\bar{\Phi}^f = \Gamma^{-1}\Upsilon,$$

where

$$\Gamma_{ij} = \int_{\mathbf{x}} \frac{a_i(\mathbf{x})a_j(\mathbf{x})}{\left[\sum_l a_l(\mathbf{x})\right]^2} d\mathbf{x},$$
$$\Upsilon_i = \int_{\mathbf{x}} \frac{a_i(\mathbf{x})f(\mathbf{x})}{\sum_l a_l(\mathbf{x})}.$$

Here l is an index over all neurons in the population.

The optimal decoders are particularly interesting in the case of narrow, symmetric tuning curves (e.g. narrow Gaussians). In the limit of narrowly-tuned neurons with centres $\mathbf{x}_i^c$, the optimal decoders are simply $\bar{\phi}_i^f = f(\mathbf{x}_i^c)$. This can be shown by discretizing $\mathbf{x}$ into a finite number of regions, so that within each region, both $f(\mathbf{x})$ and the activity of each neuron are constant. Additionally, it must be assumed for the moment that each neuron is only active in one region (modelling narrow tuning), but that many neurons can be active in a given region (modelling redundancy). In this case, if neurons i and j are tuned to different regions, then $\int_{\mathbf{x}} a_i(\mathbf{x})a_j(\mathbf{x}) = 0$. The neurons can be re-ordered so that neurons tuned to each region have adjacent indices, and the sub-matrices for each region can then be considered separately, as the off-block-diagonal entries of Γ are zero. Considering a region k, and letting i and j index neurons that are active in this region, the equation for region k can be re-written as

$$\left[\frac{a_{ik}f_k}{\sum_l a_{lk}}\Delta\mathbf{x}\right] = \left[\frac{a_{ik}}{\sum_l a_{lk}}\frac{a_{jk}}{\sum_l a_{lk}}\Delta\mathbf{x}\right]\Phi_k,$$

where $\Delta\mathbf{x}$ is the volume of region k, f_k is the uniform value of $f(\mathbf{x})$ in the k^{th} region, and a_{ik} is the uniform activity of the i^{th} neuron in the k^{th} region, and Φ_k is the vector of decoding weights for neurons active in region k. Removing common factors from these matrices, each row becomes

$$f_k = \frac{\sum_j a_{jk}\bar{\phi}_j^f}{\sum_l a_{lk}}.$$

One solution to this equation is to set $\bar{\phi}_j^f = \bar{\phi}^f$ constant for all j, so that it can be removed from the sum, and $\bar{\phi}^f = f_k$.

This simple solution for optimal weights is particularly interesting, because the weight for each synapse is independent of any information about other synapses. This contrasts with linear decoding of narrow tuning curves, in which the optimal weights for each neuron depend on how many other neurons are tuned to the same region, and on their activity levels. Independence in the average-based decoding case means that the optimal weights can be established by Hebbian plasticity. As discussed in the last chapter, this is only possible with linear decoding if the presynaptic tuning curves form a tight frame.

If tuning curves are broader (so that tuning curves with peaks in different regions overlap), but still symmetric, decoding of linear functions (i.e. $f(\mathbf{x}) = A\mathbf{x}$) is not greatly impaired (Figure 4.3). This is because while at any point $\mathbf{x_0}$, overlapping

neurons that are centred at $\mathbf{x}_0 + \delta\mathbf{x}$ will distort the estimate, neurons centred at $\mathbf{x}_0 - \delta\mathbf{x}$ will tend to compensate. However, decoding of nonlinear functions is impaired, when the nonlinearities are strong over the width of the tuning curves. Thus high-frequency functions cannot be accurately decoded by these independent weights if the tuning curves are broad.

Some computations would require a more complex projection structure than the one shown in Figure 4.2. This is because in general, the optimal synaptic weights in the direct projection might include any combination of positive and negative values, whereas a single biological neuron is most often either excitatory or inhibitory. Mixed positive and negative synaptic weights can be approximated in a realistic projection that has excitatory projection neurons and subtractive inhibitory interneurons [295] (see also Chapter 6). In the averaging code, this would lead to two groups of interneurons, one divisive and the other subtractive. In both the cortex and striatum, there are distinct groups of inhibitory neurons with preferences for different parts of the cell, which might subserve this difference. This more complex structure would not be necessary in the simple case of $\bar{\phi}^f = f_k > 0$.

Network models usually assume linear synaptic integration. It is interesting that a ubiquitous form of nonlinearity, in conjunction with a very common network architecture, can lead to an alternative decoding mechanism which has a distinct advantage for learning. A further advantage is that the average-based code is truly redundant. In a "redundant" linear-decoding network, damage to a single neuron has a small effect, but it does distort the decoding. In contrast, at least in the ideal case in which tuning curves belong to non-overlapping groups, damage to a single neuron in an averaging network does not distort the decoding at all.

But how realistic is this code? One troubling issue is that it seems to require very large synaptic weights. Individual post-synaptic potentials measured in slice preparations, where synaptic activity and associated conductances are low, are typically much too small to trigger an action potential (e.g. [386]). This is at odds with the idealized averaging scheme, in which the activity of a single presynaptic neuron would be sufficient to cause firing. However, it is not at odds with a physiological model of the averaging scheme that includes leak conductance. Leak conductance would vastly reduce the effect of any single presynaptic neuron acting alone. This would modulate the code so that it resembled a sum for few inputs, and an average for many inputs. Thus in principal, the averaging code is not inconsistent with synapses that are individually ineffective. Whether the averaging code is consistent with more realistic compartmental neuron models remains to be determined. This detailed investigation must be left for future work.

4.4 Intra-Branch Non-Linearity

An entirely different mode of decoding is supported by expansive input-current relationships within a dendritic branch.

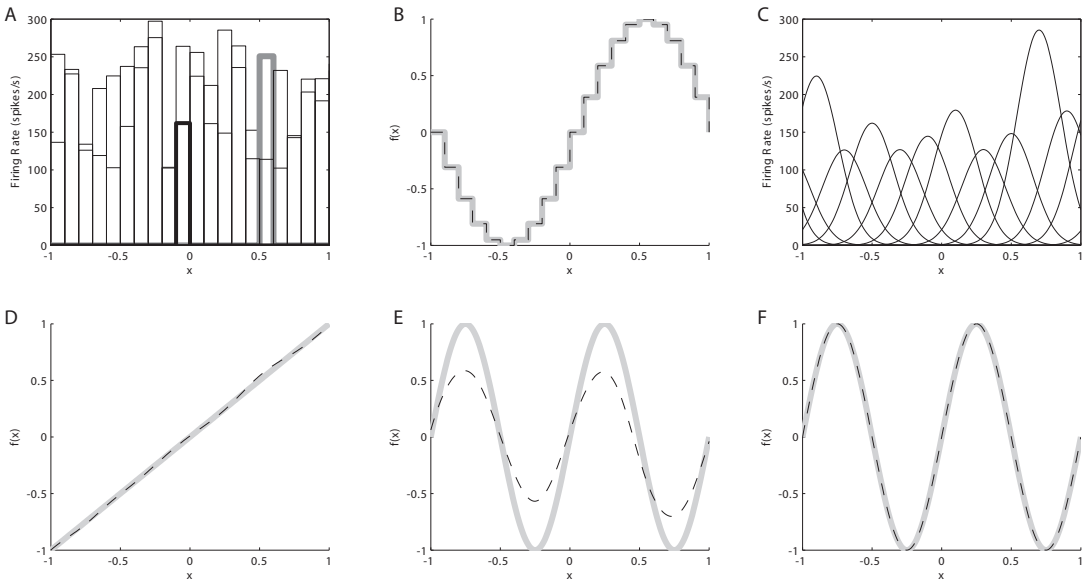

Figure 4.3: Average-based decoding. A, Narrow, flat tuning curves, which correspond to the idealized case in which each neuron's optimal decoding weight is equal to the value of the decoded function at the peak of the curve (i.e. $\bar{\phi}^f = f_k$; see text). B, An estimate of $sin(x)$ based on the tuning curves in (A), and the simple decoding weights. The gray line indicates the target function and the black dashed line indicates the decoded estimate, both at discrete values of x that correspond to the centres of the idealized tuning curves in (A). C, A small population of broader Gaussian tuning curves. D, An estimate of x, based on the tuning curves in (C), and the simple decoding weights. Despite substantial overlap between different tuning curves, and the small number of neurons, there is only mild distortion in the linear function. E, In contrast, a function in which nonlinearities are strong over the width of a tuning curve is poorly estimated by these simple decoding weights. F, The more general optimal weights (i.e. without the simplifying assumption of narrow, symmetric tuning curves) compensate for these effects, and estimate the high-frequency function well.

This type of non-linearity has been explored extensively by Bartlett Mel and colleagues. Mel [254] points out that there are several mechanisms by which the current flowing into a dendritic branch could become a supra-linear function of synaptic excitation. For example, both AMPA and NMDA receptors are gated by the neurotransmitter glutamate, but NMDA-receptor channels do not open unless the cell membrane is first depolarized from rest. Therefore in a dendritic branch with both types of channels, small increases in glutamate will open AMPA channels, while higher levels will open both types of channels, so that the glutamate-current function becomes steeper.

By fitting the input-output behaviour of various point-neuron networks to a scrupulously-detailed compartmental model, Mel and colleagues found that this type of within-branch nonlinearity can cause a dendritic tree to behave much like two layers of a sigmoidal feedforward network [302]. Each terminal branch outputs a nonlinear function of its input, and therefore resembles a single neuron in the hidden layer of the classic three-layer feedforward network. These nonlinear functions then combine linearly in the soma and pass through a spiking nonlinearity. The role of the soma is therefore analogous to that of a neuron in the output layer of the three-layer network model. This dendritic-tree model differs from a typical feedforward network model, in that a given presynaptic neuron does not synapse onto every dendritic branch (in the tree), while an input neuron is potentially connected to all hidden layer neurons (in the network). However, several factors mitigate the effect of this difference: 1) a single neuron in the brain can make multiple contacts with different dendritic branches; 2) in a large population, several neurons may have very similar tuning curves, so that they are essentially interchangeable (if at least one neuron from an interchangeable group contacts each branch, then little effective connectivity is lost); and 3) in feedforward network models, some feedforward connections have small synaptic weights, and these connections can often be pruned from the network with little effect.

It must be emphasized that this simplified model of Mel *et al.* is necessarily highly theoretical. It is very difficult to test whether such nonlinearities play a role in the behaviour of real neurons, because it is not possible to record intracellularly from terminal dendritic branches. However, if such nonlinearities do influence neuron function, they are in a position to do so profoundly. One possible effect would be to increase the memory capacity of a network [253, 303]. Such nonlinearities could also have a major impact on population coding, particularly on the functions that can be decoded from population activity, and consequently on the transformations that can be computed in a projection from one population to another.

A discussed in the previous chapter, linear synaptic integration supports transformations that are within the space of principal components of the presynaptic tuning curves. Recall that a feedforward network with a large enough hidden layer can approximate any function of its input with arbitrary accuracy. Let $\mathbf{a}(\mathbf{x})$ be the activity of a population as a function of the vector $\mathbf{x}$ that it represents. If $\mathbf{a}(\bullet)$ is invertible, then like a feedforward network, a sufficiently large nonlinear dendritic

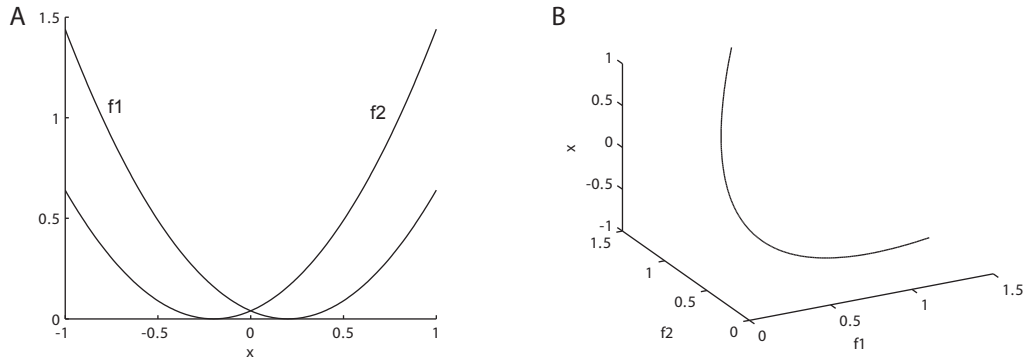

Figure 4.4: A list of functions may be invertible even of each function in the list is not. A, Two functions, $f_i(x) = (x + \alpha_i)^2$, where α_i are constants. Neither function is invertible, as its square root yields two points (for $x \neq -\alpha_i$). However, the function $f(x) = [\ f_1(x) \quad f_2(x)\ ]^T$ is invertible, because each two-dimensional point to which it maps corresponds to one value of x. B, The inverse of the function in (A), plotted for points to which this function maps.

tree can approximate any function $f(\mathbf{x})$, as

$$f(\mathbf{x}) = f(\mathbf{a}^{-1}(\mathbf{a}(\mathbf{x}))).$$

Here the dendritic tree approximates the function

$$g(\mathbf{a}) = f(\mathbf{a}^{-1}(\mathbf{a}))$$

of synaptic inputs $\mathbf{a}$ as

$$\hat{g}(\mathbf{a}) = \sum_j \sigma_j(\sum_i w_{ji} a_i),$$

where w_{ji} is the weight of the synapse from the i^{th} input onto the j^{th} dendritic branch, and σ_j is the output nonlinearity of the j^{th} branch.

Note that $\mathbf{a}(\mathbf{x})$ may be invertible even if each $a_i(\mathbf{x})$ is non-invertible, as illustrated in Figure 4.4.

The functions of $\mathbf{x}$ that can be approximated by a physical dendritic tree will be limited by the number of compartments that are effectively electronically isolated, which Mel [254] estimates to be "upwards of several dozen" in a large tree. So this type of nonlinear synaptic integration has practical limitations, but these limitations are different in character from those of linear synaptic integration.

Figure 4.5 illustrates a case in which this form of nonlinear decoding is much more versatile than linear decoding. In this example, presynaptic neurons have

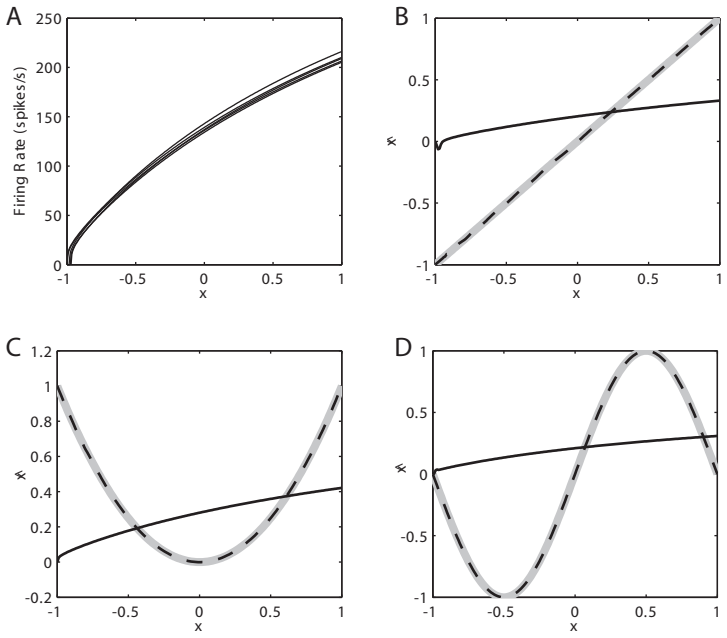

Figure 4.5: A case in which non-linear synaptic integration can closely approximate diverse functions of the input, and linear synaptic integration can not. A, Tuning curves of a group of very similar presynaptic neurons. These lack the diversity on which linear function decoding depends. B, Decoding for a communication channel. The linear target function (gray line) is closely approximated by a model of a nonlinear dendritic tree with fifty sigmoidal branches (dashed). The optimal linear decoding (solid black) does not resemble the target function. C,D The same linear and nonlinear models of synaptic integration approximating the transformations $y = x^2$, and $y = sin(\pi x)$, respectively.

highly correlated tuning curves. A very restricted set of functions can be decoded linearly from this population. In contrast, a nonlinear dendritic tree with fifty sigmoidal dendritic compartments can accurately decode a variety of functions, including the represented variable x, and other examples x^2 and $sin(x)$. This is because the branches of the nonlinear dendritic tree function much like an additional layer of neurons that is diversely tuned to the uniform output of the presynaptic neurons. Rarely are population tuning curves so uniform as in this idealized example. However, analogous situations can be encountered in population codes that are based on non-linear neuron dynamics, due to limitations in the diversity of neurons' dynamic responses (Chapter 7).

4.4.1 Invertibility of Population Responses

In order for a nonlinear dendritic tree to approximate arbitrary functions of $\mathbf{x}$, the tuning curves of the presynaptic population must be (as a group) an invertible function of $\mathbf{x}$. If they are not invertible, it will only be possible to approximate functions that have identical values, for any $\mathbf{x}$ over which the tuning curves have identical firing rates.

When are the tuning curves of a population invertible? A list of nonlinear functions $\mathbf{f}(\mathbf{x})$ is locally invertible, in some region around a point, if its Jacobian (i.e. $[\partial f_i/\partial x_j]$) is nonsingular at that point. If the functions are continuous, then the boundaries of this region extend to wherever the Jacobian becomes singular.

For cosine-tuned neurons with monotonically non-decreasing response functions (e.g. LIF; Hodgkin-Huxley), the Jacobian is

$$[\partial f_i/\partial x_j] = [\gamma_i \tilde{\phi}_i^T], \tag{4.1}$$

where $\tilde{\phi}_i$ is the preferred direction of the i^{th} neuron, and $\gamma_i \geq 0$ is the slope its tuning curve along the preferred direction. If $\mathbf{x}$ is d-dimensional, the Jacobian is nonsingular as long as there are d neurons which 1) have non-parallel preferred directions, and 2) are in regions of their tuning curves for which $\gamma_i > 0$ (for most neurons, $\gamma_i > 0$ if the neuron is active and not saturated).

4.5 Discussion

The assumption of linear synaptic integration is almost ubiquitous in neural network models. A notable exception to this rule is O'Reilly's widely-used Leabra framework [290], which incorporates nonlinear conductance-current relationships. However, while this type of nonlinearity has its greatest impact in the shunting effects of inhibition, Leabra varies inhibitory conductances in order to meet an imposed constraint on the number of active neurons, so these nonlinearities have a minimal effect on the behaviour of these models.

This chapter has discussed three types of dendritic nonlinearity that are at odds with this ubiquitous assumption. First, it was shown that when the outputs of major dendritic branches interact non-linearly, each branch can be modelled using NEF methods. The nonlinear interaction can then be interpreted as performing a high-level, nonlinear computation on the results of multiple linear decodings. If responses are diverse across a population, this type of non-linearity influences the encoding process, by shaping the tuning of each neuron to its input.

Secondly, it was shown that division due to shunting inhibition can support a variation on linear decoding, in which a post-synaptic neuron responds to weighted averages of inputs, rather than linear combinations. This decoding mechanism resembles linear decoding in some respects. However, it allows learning of transformations *via* Hebbian plasticity, in much more general circumstances.

Finally, it was shown that the theory of nonlinear synaptic integration which has been developed by Mel and colleagues supports decoding of transformations in such a way that diversity of presynaptic tuning curves is not required.

This work begins to explore the relevance of more subtle neuron properties to information processing at the population level. However, while the models in this chapter span some of the major nonlinear processes in dendrites, they do not approach the complexity of real neurons. Relating each of the intricacies of real neurons to their computational roles within larger circuits is a very long-term goal. Meanwhile, it is encouraging that the nonlinearities explored here do not require a major reworking of the NEF, but that they instead suggest useful extensions to it.

It should also be emphasized that these extensions will not always be necessary. Many successful models have assumed linear dendritic processing, suggesting that some neural systems do not rely critically on dendritic nonlinearities. Furthermore, linear models frequently lead to important insights, even when the underlying system is highly nonlinear.

4.6 Appendix: Two-Compartment Models of Division

This appendix presents two different spatial patterns of synaptic input in which inhibition has a divisive effect on firing rate.

An important parameter in these models is the extent to which spike-related shunting (i.e. a brief high-conductance phase during an action potential, which resets the membrane potential to rest) backpropagates into the dendritic tree. Hausser et al. [170] studied this parameter experimentally in pyramidal cells, and found that spike-related shunting propagated only about $200\mu m$ along the apical dendrite. This suggests that spike-shunting does not occur in the finer, more distal branches of the apical tuft. However, it is not clear how far this shunting effect reaches along the basal dendrites. More importantly, spike-related shunting may spread differently in other types of neurons. Hausser et al. also studied cerebellar Purkinje cells, and found only mild shunting, even within the soma.

In general, for other cell types, it can only be recognized that spike-related shunting may occur in the soma, and that it may spread some distance along the dendritic tree. In the first divisive regime (below), both excitation and inhibition are co-located beyond spike shunting range. In the second, both inhibition and excitation are within spike-shunting range, but excitation is distal to inhibition.

4.6.1 Distal Shunting

Both the excitatory and inhibitory neurons in this regime synapse onto dendritic regions that are too far from the soma to be influenced by spike-related shunt-

ing. The model consists of two compartments: 1) a dendritic compartment, which receives synapses, and 2) a somatic compartment, where the action potential is generated. After a spike, the membrane potential resets to its resting value in the somatic compartment, but not in the dendritic compartment.

The model is an extension of the single-compartment leaky-integrate-and-fire neuron. In a single-compartment model, with synaptic input, the membrane potential V evolves during the period between spikes as

$$\dot{V} = \frac{1}{C_m}[\sum_k g_k(E_k - V) + g_L(E_L - V)],$$

where C_m is the membrane capacitance, g_k are synaptic conductances, E_k are associated reversal potentials, g_L is the leak conductance, and E_L is the reversal potential of the leak current, which equals the cell's resting potential (i.e. the potential to which it decays without synaptic input). When the membrane potential exceeds the spike threshold, it is reset to its resting value. Then, after a short refractory period, integration of the input begins anew.

This model can be extended to multiple compartments by adding terms $g_{ab}(V_a - V_b)$, to model the current flowing from compartment a into compartment b [92]. The membrane capacitance and conductances in the equation above are per unit area of membrane. Consequently, the current flowing between compartments must be normalized by the relative membrane surface area of each compartment, so that generally $g_{ab} \neq g_{ba}$.

In the model of the distal divisive regime, the membrane potential in each compartment evolves as

$$\begin{aligned}
\dot{V_1} =& \frac{1}{C_m}[g_i(E_L - V_1) + g_e(E_e - V_1) + g_{21}(V_2 - V_1) + g_L(E_L - V_1)], \\
\dot{V_2} =& \frac{1}{C_m}[g_{12}(V_1 - V_2) + g_L(E_L - V_2)],
\end{aligned}$$

where V_1 and V_2 are the membrane potentials of the dendritic and somatic compartments, g_i is inhibitory condutance, g_e is excitatory conductance, and E_i and E_e are the corresponding reversal potentials. Note that all the synaptic input is onto compartment 1 (the dendritic compartment), and that the reversal potential of inhibitory synapses is equal to the resting potential. The dendritic compartment is taken to have a much larger surface area than the somatic compartment (e.g. in pyramidal cells, dendritic surface area is one or two orders of magnitude greater), so that $g_{12} >> g_{21}$.

Figure 6A shows a simulation of this model with constant excitatory input. Inhibition doubles half way through the simulation, causing the firing rate to drop by a factor of two. There is a startup transient, in that the time to first spike is longer than the subsequent inter-spike intervals. This is because $V_1(0) = V_2(0) = E_L$, whereas only V_2 is reset to the resting potential after each spike.

Recall that in the single-compartment model, inhibition has a subtractive effect on firing rate, because its divisive effect on membrane potential is cancelled out by its divisive effect on the membrane time constant. The situation is the same in the distal compartment of this model. However, V_1 is constant after the first spike, so that the time constant of the membrane in this region is of no consequence. In contrast, the equilibrium value of V_1 is affected by inhibition. V_1 determines the rate of current flow into the somatic compartment, which in turn determines the spike rate.

4.6.2 Proximal Shunting

This model also has two compartments: 1) a dendritic compartment that receives excitatory input, and 2) a somatic compartment that receives inhibitory input. In contrast with the distal shunting model (above), the membrane potential in this model resets in both compartments after each spike. This means that the inhibitory conductance in compartment (2) affects the model's sub-threshold dynamics. However, the dynamics are dominated by a slower membrane time constant in compartment (1).

Between spikes, the membrane potential in each compartment evolves as

$$\begin{aligned}\dot{V_1} =& \frac{1}{C_m}[g_e(E_e - V_1) + g_{21}(V_2 - V_1) + g_L(E_L - V_1)],\\ \dot{V_2} =& \frac{1}{C_m}[g_i(E_L - V_2) + g_{12}(V_1 - V_2) + g_L(E_L - V_2)].\end{aligned}$$

The membrane time constants τ of the two compartments are

$$\begin{aligned}\tau_1 =& \frac{C_m}{g_L + g_e + g_{21}},\\ \tau_2 =& \frac{C_m}{g_L + g_i + g_{12}}.\end{aligned}$$

The leak conductance g_L is small compared to the synaptic conductances [102]. So, since $g_i > g_e$ and $g_{12} >> g_{21}$, the time constant τ_1 is much larger (slower) than τ_2. This means that when excitatory input arrives at compartment (1), most of the delay in the build-up of somatic membrane potential is due to compartment (1). Compartment 2 responds quickly and introduces little extra delay in the spike. Since inhibition only affects this minor delay, it has little effect on the dynamics. However, inhibition still has the same divisive effect on voltage that it had in the single-compartment model, so that the net effect on the firing rate is divisive.

Figure 4.6 shows example simulations and tuning curves both the distal and proximal shunting models.

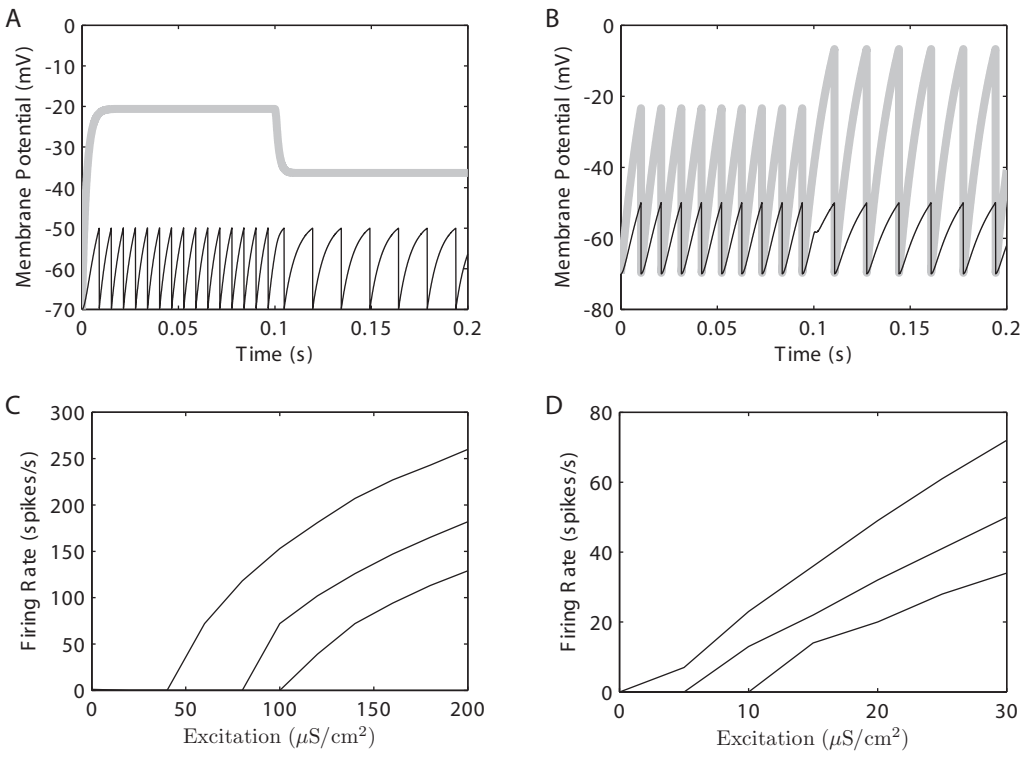

Figure 4.6: Distal (A&C) and proximal (B&D) shunting models. The top panels show example simulations in which excitation is held constant, and inhibition is doubled at 0.1s. The gray line indicates the membrane potential in the dendritic compartment, and the black line indicates the membrane potential in the somatic compartment. This simulation includes the sub-threshold regime only. A spike (not shown) is generated when membrane potential in the soma reaches -50mV. The bottom panels show the tuning curve of a single neuron of each type, as a function of excitatory conductance. The separate lines indicate three different levels of inhibition. In each case, inhibition has an effect similar to stretching the excitation axis: the slope of the tuning curve is reduced, and firing threshold is increased.

Chapter 5

Temporal Coding

Fine temporal patterns of firing in the basal ganglia, as in much of the brain, are highly irregular. In some circuits, the precise pattern of irregularity contains information beyond that contained in mean firing rates. However, the capacity of neural circuits to use this additional information for computational purposes is not well understood. This chapter shows that an ensemble of neurons firing at a constant mean rate can induce arbitrarily-chosen temporal current patterns in postsynaptic cells. If the presynaptic neurons fire with nearly uniform interspike intervals, then current patterns are sensitive to variations in spike timing. But irregular, Poisson-like firing can drive current patterns robustly, even if spike timing varies by tens of milliseconds from trial to trial. Notably, irregular firing patterns can drive useful patterns of current even if they are so variable that several hundred repeated experimental trials would be needed to distinguish them from random firing. Together, these results describe an unrestrictive set of conditions in which postsynaptic cells might exploit virtually any information contained in spike timing.

5.1 Introduction

Past theoretical and experimental work has shown how inter-neuronal communication through firing rates supports a wide range of computational processes (Chapter 3). However, in some systems, additional information is contained in the precise timing of action potentials (e.g. [288, 389]). Information-theoretical studies have extensively characterized the amount of information carried by action potential timing in sensory systems [314]. Although less widely studied, timing also appears to be important in motor and frontal areas [6, 313]. There are also several reasons to think that spike patterns are important in the basal ganglia. Spike patterns in basal ganglia nuclei are abnormal in Parkinsonism, and these abnormalities seem as prominent as abnormalities of mean firing rate. Spike patterns change with symptomatic improvement, on administration of anti-Parkinsonian medication, sometimes without accompanying changes in firing rates [225]. If symptoms arise partly

from abnormal patterning, this might explain why both lesion and deep brain stimulation of the GPi or STN have similar effects [352], and it would be easier to account for the improvement of dyskinesias with GPi lesion. Also, while the Albin/DeLong model predicts a decrease in GPi activity with Huntington's disease, and an increase with PD, one recent study [356] (although the sample size was small) reported that the mean rates in these diseases were very similar, but that patterns of spiking were not.

The functional relevance of any information contained in spike timing depends entirely on what postsynaptic neurons can do with this information. This motivates the focus in this chapter on the effects that timing-based information can have on postsynaptic cells. It is well-established that action potential timing plays a role in synaptic plasticity (see reviews by [197, 87]), but spike timing can also underlie computational processes. For example, activity in a neuron can depend on the degree of synchrony between the presynaptic neurons that converge onto it [5, 345, 340, 324]. This phenomenon underlies perception of the horizontal location of low-frequency sound sources [398, 56] and has been suggested to play a significant role in high-level visual perception (although see [337, 86]) and the recognition of odors [234, 58]. Notably, synchrony-based computations can also be performed with asynchronously generated spikes, provided propagation times differ so that spikes arrive synchronously at their target [161, 274, 177].

Less is known about how the timing of action potentials can affect computational processes in the absence of synchrony. But a number of cases demonstrate that the effects can be substantial. For example, information about tactile stimuli that are applied to human fingertips is encoded in the relative timing of the first spikes from different sensory neurons [187]. This information can be extracted effectively by a projection with unequal excitatory synaptic weights and parallel inhibition [361]. Similarly, information contained in the timing of consecutive spikes (in one neuron) can be extracted by certain types of synapses [273], neurons [334], or specific circuits [7, 61, 205]. Also, some learning rules can lead simple neuron models to support a wide variety of mappings between incoming spike patterns and output [218, 142]. These examples illustrate that in a variety of situations, postsynaptic neurons may read out information contained in spike timing without relying on synchrony. However, the relevance of nonsynchronous spike timing to the operation of neural circuits in general remains uncertain.

In particular, it is not yet clear whether nonspecialized neurons can use information encoded in arbitrary spike patterns in a flexible manner, that is, to compute arbitrary functions of the encoded signals. In this direction, Legenstein et al. [218] have shown that spike-timing-dependent plasticity can lead to input/output mappings that correspond to arbitrarily chosen sets of synaptic weights, but this does not clarify whether mappings to arbitrarily chosen output spike patterns are possible. As discussed below, the latter question has important implications for the interpretation of electrophysiological data. Therefore, this chapter addresses the question of whether there exist sets of synaptic weights that will transform arbitrarily selected patterns of spike timing into arbitrarily selected temporal patterns

of current in a post-synaptic neuron model.

To answer this question, a conductance model is used to characterize synaptic currents, adjusting weights so that synaptically induced current at the soma optimally approximates preselected target patterns. It is shown that commonly observed types of firing patterns can drive a wide variety of current patterns in postsynaptic cells, regardless of whether their mean rates vary over time. This remains true even if spike times vary randomly with a standard deviation (SD) of more than 10 ms. In some cases, effective postsynaptic currents (PSCs) can be driven by firing patterns that are so variable that the probability of distinguishing them from random firing is remote. Thus, in very general circumstances, the information contained in patterns of spike timing can be read out as arbitrary patterns of current in a postsynaptic cell. The chapter concludes by suggesting how this phenomenon may underlie a versatile population-temporal coding scheme.

5.2 Methods

5.2.1 Approximation of Current Patterns

The key procedure in this chapter is the assessment of how well given firing patterns can induce preselected patterns of current in a postsynaptic cell model. As discussed below, the target current was never induced exactly, but for a given presynaptic firing pattern, approximations of varying quality could be obtained by adjusting synaptic weights. The approximation of interest was the best one that could be obtained for each firing pattern/ target current pair.

Target currents were approximated by a linear combination of the PSCs that were induced at each synapse in a model cell. The optimal synaptic weights for approximating a given target current were found by adapting the NEF method for decoding neural representations of scalars [111]. The following error function was minimized (using the Moore-Penrose pseudoinverse) with respect to synaptic weights w:

$$E = \int^{T} [I(t) - \sum w_i I_i]^2 dt,$$

where E is the error, $I(t)$ is the current pattern to be approximated, w_i is the weight of the i^{th} synapse, I_i is the unweighted PSC pattern at each synapse, and t is time. In cases where firing patterns varied from trial to trial due to noise, the above integral was evaluated over 32 repeated trials to find optimal weights, and performance was then evaluated as the average mean-squared error (MSE) over 5 additional trials. Accuracy improved with greater numbers of trials, but improved little with 64 as opposed to 32 trials.

The model of current dynamics at each synapse was adapted from a model of alpha-amino-3-hydroxy-5-methyl-4-isoxazoleproprionic acid (AMPA) receptors [101]. This model determined the temporal shape of the current at each synapse,

whereas the optimal synaptic weights determined the absolute scale. The results were not sensitive to alternate PSC models, different time constants of current decay, or diverse time constants at different synapses. The common simplifying assumption was adopted that synaptic currents combine linearly at the soma (e.g., [142, 177]). As discussed in Chapter 4, this is a reasonable approximation of some, but certainly not all, cases of synaptic integration. Linear combination was achieved by holding membrane potential (at the synapse) at –65 mV, a constant far from the reversal potential. By summing conductances instead of currents, the analysis can be generalized to any case in which there is a monotonic relationship between conductance and current, but this additional complexity is avoided here.

The study focused on target current patterns in the 0- to 5-Hz band, which approximates the range of frequencies over which neural firing rates change in many circuits. For example, muscle activation patterns in humans (which are rate coded) consist mainly of frequencies under 5 Hz. A selection of band-limited target currents was generated by assigning random coefficients to different frequency components and calculating the inverse fast Fourier transforms.

5.2.2 Presynaptic Firing Patterns

Presynaptic firing patterns were obtained in 2 different ways. First, an initial study was performed with firing patterns produced by a cortical network model. Second, synthetic spike trains with desired statistical features were generated by drawing interspike intervals (ISIs) from appropriate probability distributions. These methods are described in detail below.

Network Simulation

The cortical network model [176] consisted of 200 fast-spiking inhibitory and 800 excitatory neurons, the latter mainly adapting with some bursting neurons. In some simulations, the coefficient of variation (CV; i.e., the SD divided by the mean) of ISIs (within the spike train of each neuron) was increased. CV was increased by shifting the excitatory neuron distribution to favor bursting neurons and decreasing excitatory coupling by 40%.

Synthetic Spike Trains

Synthetic spike trains were used to explore in detail how the results obtained from the cortical network model related to its patterns of firing. ISIs for synthetic spike trains were drawn from 3 types of probability distributions: Gaussian centered on a mean firing rate (repetitive spiking), a shifted exponential distribution with zero probability between 0 and 2 ms (Poisson-like pattern with refractory period) and a bimodal distribution consisting of the sum of 2 Gaussians, chosen so as to obtain a specified mean rate and CV = 2 (irregular bursting). To obtain spike trains with

CV < 1, the Gaussian and exponential distributions were combined in a weighted average. Spike trains with CV between 1 and 2 were obtained by averaging the exponential and bimodal distributions.

Because each synthetic firing pattern was generated from a single ISI distribution, these patterns are referred to for present purposes as having constant firing rates. Because the mean rates do not change over time, ISI ordering makes up all the information content of these firing patterns. This means that, for example, the Poisson patterns in this chapter are not treated as Poisson noise, but as information with Poisson statistics. Noise was introduced separately, either as spike time jitter or in the form of additional spikes that were introduced at random from trial to trial.

It was hypothesized that firing time correlations across different neurons might also affect performance, separately from the effects of the temporal regularity of firing patterns. Spike trains with different levels of pairwise correlation were produced in 2 ways.

Method A: Spikes were distributed in a Gaussian pattern (SD = 3 ms) around Poisson-distributed correlation times [49]. The degree of correlation was varied by changing the rate of correlation times relative to the firing rate. For example, when the rates were similar, each spike train contained a spike at almost every correlation time, and pairwise correlations were very high. Correlations were low when the firing rate was much lower than the rate of correlation times.

Method B: Poisson firing rates R in each spike train were varied over time according to the template function: $R = A[sin(2\pi Bt) - C]^+$, where $[]^+$ indicates positive rectification, B =10, 22, or 55Hz, t is time, C is a threshold between –2.0 and 0.9, and A is a constant that normalizes the template to produce the desired mean firing rate. At higher thresholds, firing only occurred at peaks of the sine wave, resulting in high correlations.

The index of pairwise correlation reported here is the peak cross correlation

$$R = (R_{AB} - N_A N_B/N)/[(N_A - N_A^2/N)(N_B - N_B^2/N)]^{1/2},$$

where R_{AB} is the number of coincidences in each 1-ms bin, N_A and N_B are the numbers of times that cells A and B fire, and N is the number of bins (e.g. [364]). These methods result in similar degrees of correlation between different pairs in an ensemble. This is a simplification, in that there is typically substantial variation between pairwise correlations in a real neural ensemble.

5.2.3 Statistical Power Analyses

Statistical power analyses were performed in order to determine the numbers of experimental trials that would be needed to detect the subtlest firing patterns that could drive reproducible activity in postsynaptic targets (see details in the chapter appendix). These analyses apply to experiments that consist of repeated

recordings of a single cell from a population with Poisson firing statistics. Cells that are postsynaptic to this population may also receive inputs from other populations, but the net effect of other inputs is assumed to be nearly constant.

5.3 Results

5.3.1 Cortical Network Simulation

A simulated network of 1000 irregularly firing cortical neurons [176] was able to generate PSCs that closely approximated a wide variety of target patterns. Figure 5.1 shows current patterns generated simultaneously by this network in 3 different postsynaptic cell models, which differ only in terms of synaptic weights. The current pattern in the first cell is a smoothed and scaled version of the network's mean firing rate. This is the type of current pattern that would emerge with uniform or random synaptic weights, so it is not surprising that this target pattern can be approximated very closely when synaptic weights are optimized specifically for this purpose. The current pattern in the second cell is an arbitrarily chosen square pulse. In contrast with the current pattern of the first cell, this current pattern is not related to the network's firing rate or to any other time-varying statistic of the network's activity. However, with appropriately chosen synaptic weights, this pattern is also approximated accurately. The current pattern in the third cell consists of randomly selected frequency components in the 0- to 5-Hz band. Like the square pulse, it bears no obvious relationship with the network's firing pattern, but it is also well approximated. Somatic current in each of these cells deviates less than 1% from the target, in the mean-squared sense. These examples show that a given pattern of firing may drive an extremely wide variety of PSCs given appropriately-chosen synaptic weights.

5.3.2 Firing Pattern Regularity

This basic result does not address how statistical features of a population firing pattern might constrain the current that it can induce in a postsynaptic cell. Synthetic spike trains were used to explore this question in detail. Approximation error was found to depend strongly on the regularity of spike trains over time. Figure 5.2 (panels A-D) shows approximations of band-limited current patterns by firing patterns with differing temporal regularity. Notably, spike trains with essentially-constant firing rates (e.g., Fig. 5.2A,B) could approximate arbitrarily chosen time-varying current patterns in the postsynaptic cell model. However, error was markedly reduced as the CV of ISIs increased.

These results are not surprising when the currents at individual synapses are considered in the frequency domain. The currents at individual synapses can be

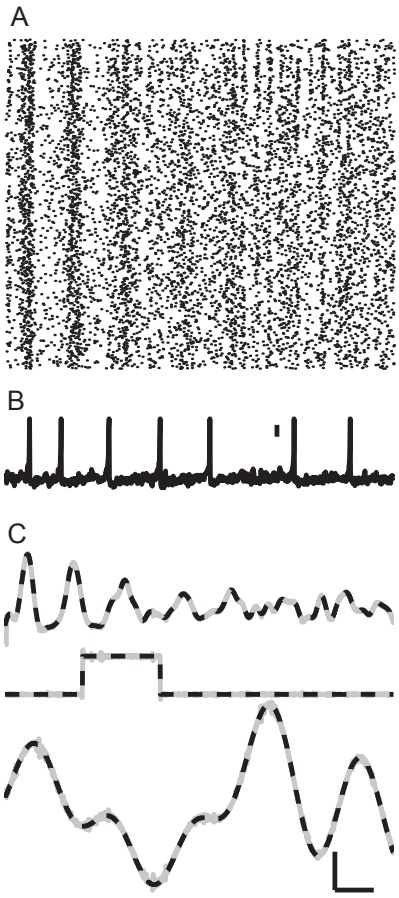

Figure 5.1: Pattern generation example. A model of 1000 cortical neurons [176] can generate arbitrarily chosen current patterns in a postsynaptic cell. A, Spike times (one neuron per row). B, Membrane potential of a typical excitatory neuron in this network (scale bar 20 mV). C, Current induced in 3 different postsynaptic cells, to which the network projects with different synaptic weights. Currents are optimal approximations (gray) of target patterns (black dashed). Top: smoothed and scaled reflection of the network's mean firing rate, middle: an arbitrarily chosen square current pulse, bottom: an arbitrarily chosen band-limited target (scale bars: 1 nA and 100 ms). Timescale in (C) applies to all panels.

viewed as temporal basis functions, which are weighted and summed to approximate the target pattern. The frequency content of these basis functions depends on the firing pattern of the corresponding presynaptic cell. For example, the current that arises from regular firing consists of harmonics of the firing frequency, whereas that arising from Poisson firing has a broad spectrum. This can be seen in the lower traces of Figure 5.2A-D, which show the power spectra of the first several principal components of the PSCs that are induced by each ensemble. Approximation error decreases with increasing power in the frequency range of the target current and increases with increasing power at other frequencies. As a result, both Poisson-refractory and irregular-burst firing patterns can accurately generate target currents with a wide range of frequencies. Burst firing is more effective than Poisson-refractory firing for driving low-frequency current patterns. However, Poisson-refractory firing is effective over a slightly wider frequency range (Fig. 5.2E).

Firing patterns in most neural circuits tend to have high CV. These results begin to suggest that information contained in such patterns can be extracted in an accurate and flexible manner.

5.3.3 Spike Jitter and Noise Spikes

The results described so far are highly idealized in that they are based on noise-free firing patterns. In order to quantify the dependence of current generation accuracy on precise spike timing, simulations with synthetic spike trains of different CV were repeated with random (Gaussian distributed) spike time jitter.

Spike jitter with a given variance had the effect of increasing MSE by a near-constant multiple, regardless of CV. Thus at high CV, where error without spike jitter was minimal, error remained relatively low even when substantial jitter was applied. For example, with bursting spike trains (CV > 1), 8-ms jitter resulted in error of at most 5% of root-mean-squared current (Figure 5.2F). Similar results were obtained when firing patterns were corrupted by inserting additional "noise spikes," at random times (determined by a constant-rate Poisson-refractory process) that were uncorrelated between repeated trials (Figure 5.2G).

Figure 5.3 shows an example in which half of the spikes are noise spikes and the other half are subject to extreme Gaussian jitter ($\sigma = 20$ ms). The target pattern is nevertheless approximated with reasonable accuracy, illustrating that meaningful population output requires very little consistency in the fine temporal firing patterns of individual neurons, even in the absence of coarse firing rate variations.

5.3.4 Population Size and Firing Rate

For firing patterns with a given CV, error decreased with increasing presynaptic population size (Figure 5.4). However, unrealistically-large populations were not

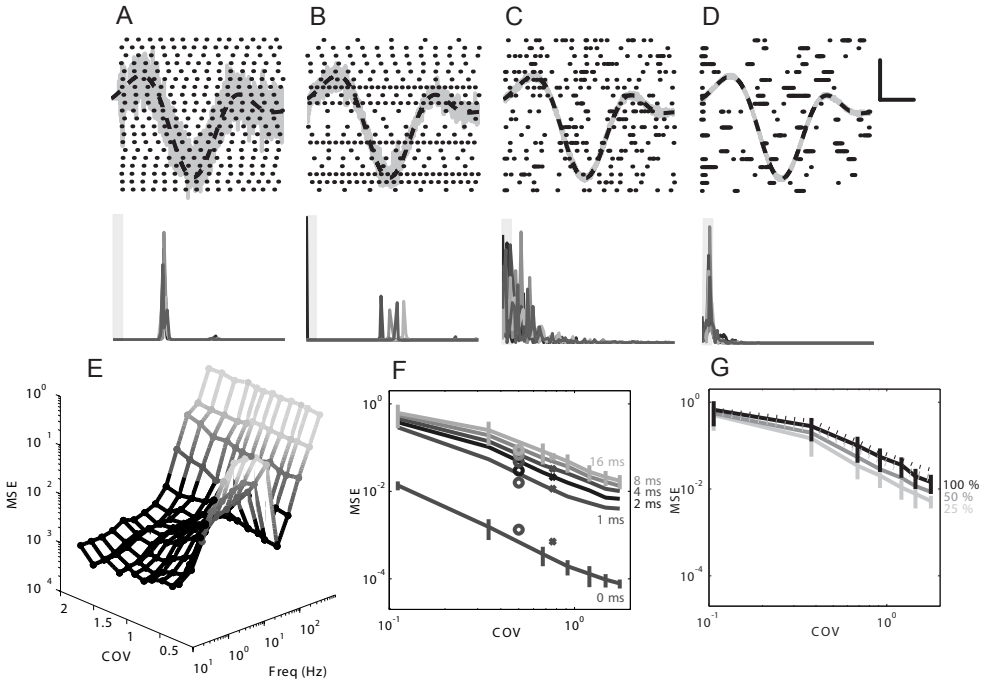

Figure 5.2: Decreasing error with decreasing spike pattern regularity. All data are from simulations with 500 synthetic neurons, with mean firing rate 30 Hz, but different ISI distributions. In panels A-D, dots indicate spike times of example neurons, black dashed lines are target currents, and gray lines are actual net synaptic currents flowing into the post-synaptic cell model. Traces below are power spectra of the first 5 principal components of the PSCs (range 0-100 Hz; shaded area 0-5 Hz). A, Neurons that fire at near-constant rates (CV=0.08; MSE=0.117 nA). B, Constant rates with wider rate distribution (across neurons) than in (A) (CV=0; MSE=0.015 nA). C, Poisson-refractory neurons (CV=0.94; MSE=0.002 nA). D, Irregular-bursting neurons (CV=1.7; MSE=0.0003 nA). E, MSE (as a proportion of root-mean-squared target current amplitude) in approximating sinusoids of different frequencies (mean over 5 different phases at each frequency) for a wide range of CV. Error is generally high with low CV, except when sinusoid frequency is close to the mean firing frequency. F, MSE versus CV. Separate lines are degrees of Gaussian jitter (SD as labeled). Error bars on top and bottom traces indicate SD over 5 randomly selected band-limited signals. Symbols O and X indicate means for a 500-neuron version of cortical network and for the same network adjusted for higher CV (see Methods), respectively. G, As (F) but with noise in the form of additional, randomly timed spikes instead of jitter. Number of noise spikes given as percentage of number of non-noise spikes. Dashed lines of the same shade indicate errors with the same proportion of noise spikes plus 4-ms jitter

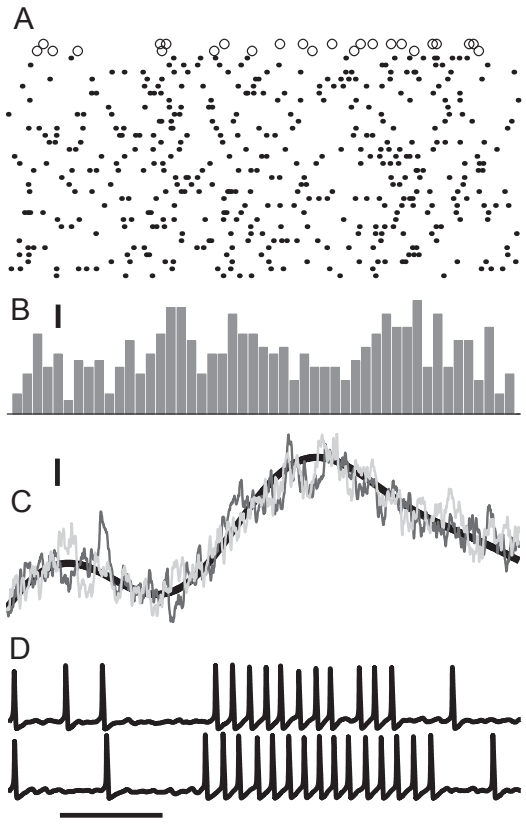

Figure 5.3: Moderate error with highly variable spike trains. The presynaptic population consists of 1500 synthetic Poisson-refractory spike trains. Each train consists of 2 interlaced 20 spike/s components. One component is subject to large spike jitter (SD=20 ms) that is uncorrelated between trials. The other component is completely uncorrelated between trials (i.e., in each trial, this component consists of a new set of spikes from a Poisson-refractory process, which is independent of previous sets). A, Spike times of an example presynaptic neuron, over 32 trials used to find synaptic weights (dots), and 2 separate trials shown in panel (C) (circles). B, Spike time histogram of a single example neuron (scale bar: 10 spikes/s). C, Approximations (gray) of target current (black) for the 2 trials shown as circles in (A) (scale bar: 2 nA). D, Membrane potential of a Hodgkin-Huxley model [207] driven by the 2 current approximations shown in (C) (scale bar: 100 ms applies to all panels).

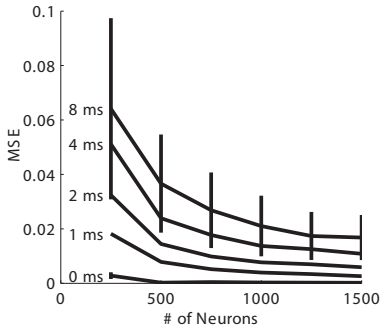

Figure 5.4: Error decreases with increasing population size. Results from Poisson-refractory neurons (40 spikes/s), with different degrees of Gaussian spike time jitter are shown (jitter SD as labeled). Error bars on top and bottom traces indicate mean ± SD of MSE over 5 randomly selected band-limited target currents (as a proportion of root-mean-squared target current amplitude). Error varies with spike jitter as in Figure 5.2.

needed. For example, with 1-ms spike jitter, 1000 presynaptic Poisson-refractory neurons were adequate to generate 500-ms signals with roughly 2% MSE.

In contrast with population size, firing rate had little effect on the accuracy of current generation. Errors arising from Poisson-refractory inputs were consistent over a wide range of intermediate firing rates, increasing slightly both below 5 spikes/s and above 100 spikes/s (Figure 5.5). Thus errors were low over a wide physiological range. The increase in error with higher rates is related to the fact that the refractory time causes a more pronounced deviation from Poisson statistics (lower CV) at higher rates. This can be seen by comparing the solid and dashed lines in Figure 5.5.

5.3.5 Correlated Firing

This chapter so far has essentially characterized the synaptic currents that arise from irregular firing as having low-frequency components that form an overcomplete temporal basis of possible somatic currents, over some range of frequency and time. Because such functions span a larger space if they are linearly independent, it was hypothesized that spike timing correlations would impair performance. Synthetic spike trains were used to test this prediction (this is separate from the question of correlated noise, studied by e.g., [2, 327]). Error generally increased with correlated spike timing. This is because when spikes were concentrated around correlation times, there were fewer spikes in the intervening periods, which is analogous to the population briefly consisting of fewer neurons (see previous section). However, the increase in error was minimal when correlation times were periodic at high frequencies (Figure 5.6). This can be explained by noting that when correlation

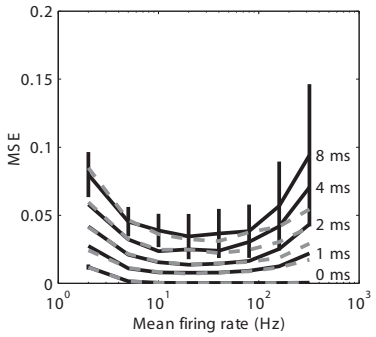

Figure 5.5: Error is nearly constant over a broad range of firing rates. Separate lines correspond to Gaussian jitter with SD as labeled. Solid black: Poisson-refractory neurons. Dashed gray: Poisson neurons. Error bars on top and bottom traces indicate mean ± SD of MSE over 5 randomly selected band-limited target currents (as a proportion of root-mean-squared target current amplitude).

times are frequent, some of the PSCs that begin flowing around one correlation time will continue to flow until the next, so that the effective population size remains large throughout. These results suggest that although correlated firing may underlie some forms of temporal coding, it may preclude other forms that rely on diverse timing to support a wide range of temporal transformations.

5.3.6 Learning

The results presented above are based on synaptic weights that were obtained using an artificial optimization method. The physiological relevance of these results depends on whether each synaptic weight can be independently learned, using only information that is available at the corresponding synapse. This section shows that synaptic weights can indeed be learned in this manner, provided some explicit error or target signal is available.

The derivative of the error function defined earlier (see Methods), with respect to each synaptic weight, equals the inner product of the current and the error over time. This suggests a supervised learning rule in which each synaptic weight is updated at each instant, by $\Delta w_i = -\kappa I_i^{syn} E$, where κ is a constant learning rate, I_i^{syn} is the instantaneous current at the i^{th} synapse, and E is the instantaneous error in net current. This learning rule quickly converges on results similar to those obtained with the optimization method (Figure 5.7). This remains true in the presence of spike jitter.

Assuming an error signal were available, it is doubtful whether this signal would propagate instantly to each synapse. The performance of the learning rule was therefore investigated when Δw_i was based on low-pass filtered error and current

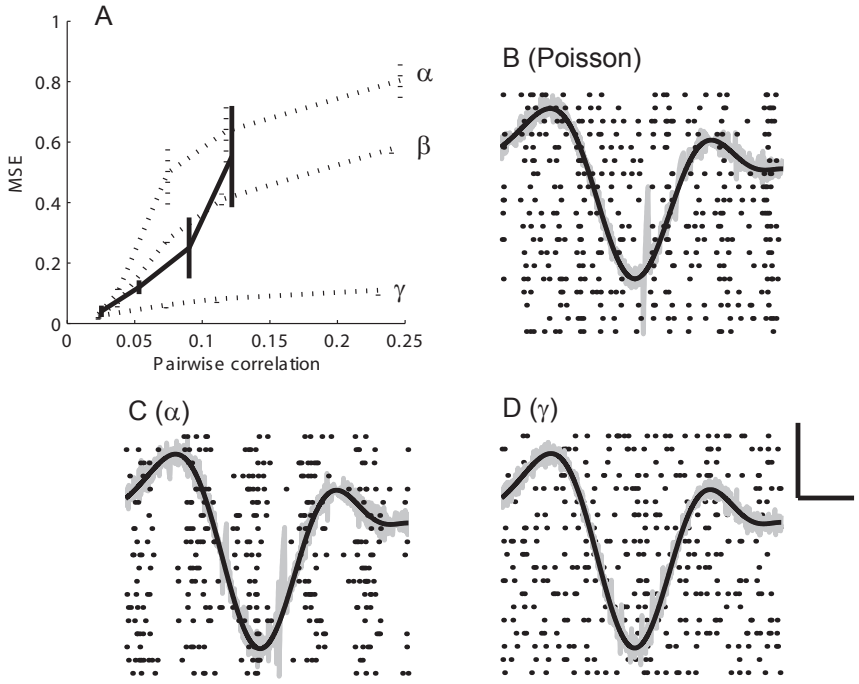

Figure 5.6: Increasing error with increasing spike time correlation. A, MSE versus correlation (4-ms spike jitter) with 500 Poisson-refractory neurons (40 spikes/s). Solid and dashed lines indicate Poisson and periodic correlation times, respectively (see Methods; α =10Hz; β =22Hz; γ =55Hz). MSE reported as proportion of root-mean-squared target current amplitude; bars indicate SD over five 300-ms targets. B-D, Examples of approximations with Poisson, α, and γ correlations of roughly equal strength. Dots indicate spike times of example neurons, black lines are target currents, and gray lines are the actual synaptic currents flowing into the postsynaptic cell model. Scale bars: 100 ms and 1 nA.

signals. Filtering obscured high-frequency errors from the learning mechanism. Consequently, learning was slowed, and the resulting approximations contained more noise in the frequency range corresponding to the stop band of the filter (Figure 5.7). However, these limitations were not severe. Reasonable approximations were obtained even when the filter time constant was greater than the duration of the target signal. This demonstrates that learning can proceed on the basis of error information that is substantially lagged and temporally smoothed, converging nonetheless close to the limit imposed by spike jitter.

5.3.7 Experimental Detection of Subtle Repeated Patterns

As previously demonstrated, spike patterns with little trial-to-trial consistency can drive highly consistent activity in a postsynaptic target (Figure 5.3). This raises the question of whether spike patterns that have a stereotyped relationship with behaviour might be driven by spike patterns that are so variable with respect to behaviour that any underlying consistency evades experimental detection. Statistical power analyses were performed to address this question. The analyses estimate the numbers of repeated trials that would be needed to find peri-event variations in firing rate, under the assumption that such variations are as small as possible while still producing relatively reliable spiking in a postsynaptic cell.

Figure 5.8 shows the numbers of trials that would be needed to detect the subtlest presynaptic firing patterns that could drive post-synaptic firing with various levels of consistency. The number of trials needed (recording a single representative presynaptic neuron) depends strongly on how reliable *post*-synaptic spiking is assumed to be. This is because the more pronounced variations in presynaptic firing that would be needed to cause more reliable post-synaptic firing would also require fewer trials to detect. However, even if post-synaptic spiking were highly stereotyped (1% of spikes timed inconsistently from trial to trial), 50 or more repeated trials may be needed to distinguish the driving patterns from random firing. Throughout the range of error rates shown in Figure 5.8A, trial-to-trial consistency is greater in postsynaptic than in presynaptic firing patterns. So, presynaptic firing patterns that are so subtle as to require over 1000 trials to detect may nevertheless drive much more stereotyped activity in postsynaptic cells. Although the specific results of this analysis clearly depend on the assumptions made (e.g., degree of convergence; Poisson firing statistics), similar assumptions are reasonable with respect to many cortical and subcortical areas. The key observation is that substantially more trials may be needed to detect useful repeated firing patterns (e.g., over 100 trials, if a 10% rate of postsynaptic spike mistiming is assumed) than are typically collected in experimental studies (except in studies in which repeated trials consist only of brief sensory stimuli, e.g., [33]).

In fact, these results may underestimate the capacity for highly variable spiking to produce stereotyped behaviour, because the power analyses ignore potential dynamic effects. Specifically, firing at the output of a network will have greater

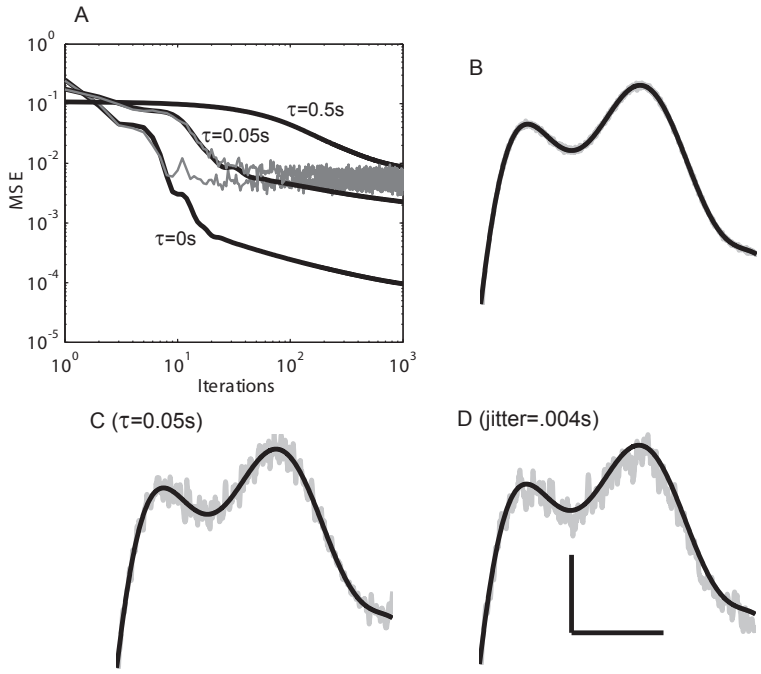

Figure 5.7: Learning. A, Decrease in error over 1000 iterations of a Poisson-refractory spike pattern (500 neurons; 30 spikes/s), under the learning rule described in the text. All synaptic weights initially set to zero; target current as shown in other panels. Thick black lines indicate learning trials with no spike jitter. Three cases are shown, each with error data temporally filtered using a different first-order low-pass filter (time constants as labeled; $\tau = 0s$ indicates no filter). The thin gray lines that diverge from the black lines after about ten iterations indicate corresponding cases repeated with 4 ms (SD) jitter in the spike trains (only the $\tau = 0s$ and $\tau = 0.05s$ cases are shown). Interestingly, there were substantial differences in error after a single iteration (left extreme of each line), depending on the filter time constant. Substantial filtering allowed the learning mechanism to accurately approximate the mean magnitude of the target signal in a single pass, although subsequent learning of the signal shape was slowed. Learning continued after 1000 iterations (not shown). For example, with $\tau = 0.5$, error was further reduced by about half, after 10 000 as opposed to 1000 iterations. Panels B-D show target current (black) and approximation (gray) in various cases, after 1000 iterations. B, Neither filter nor spike jitter. C, Filter with $\tau = 0.05s$. D, Spike jitter with SD = 4 ms. Scale bars: 100 ms and 0.5 nA.

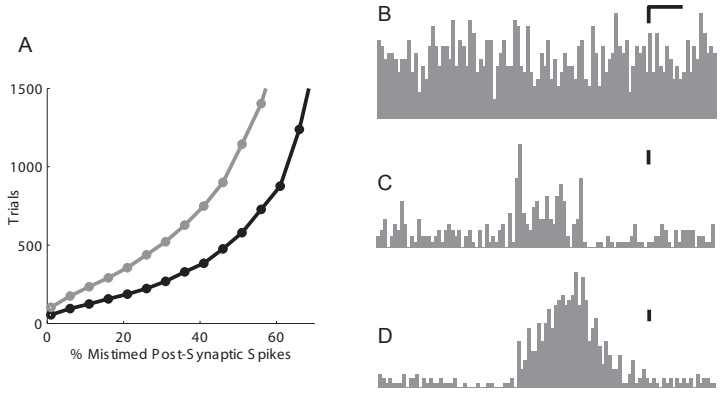

Figure 5.8: Trials needed to detect subtle firing patterns. Results of prospective power analyses for (hypothetical) experiments to detect the smallest peri-event firing rate changes that could trigger reliably-timed spiking in a post-synaptic cell. Assumptions are as described in the Methods. Details of the analysis are given in the Appendix. A, Numbers of trials required for a type-II error rate of 0.2 with 1-way ANOVA. More trials are needed to detect smaller presynaptic fluctuations in firing rate. The expected size of presynaptic rate fluctuations depends on the number of neurons contributing to each post-synaptic spike (black: 500; gray: 1000) out of a total of 10 000, and on the reliability with which the post-synaptic cell is assumed to spike. For example, larger presynaptic variations in firing rate lead to more reliable post-synaptic timing and also require fewer trials to detect. An impractically large number of trials may be needed to detect subtle patterns, unless it can be assumed that the patterns drive post-synaptic activity with a very low error rate. B, 100-trial spike timing histogram for an example neuron drawn from a population that drives post-synaptic firing with a mistimed spike rate of about 60%. C, 100-trial firing histogram for a Hodgkin-Huxley neuron driven by the population exemplified in (B) with PSC time constant of 5 ms. D, As (C) but with 20 ms PSC time constant.

consistency if the network is more responsive to underlying firing patterns than to random fluctuations. Figure 5.8D shows the results of a simulation that illustrates this point using a Hodgkin-Huxley model [207] of a post-synaptic neuron. In this simulation, the receiving neuron is made less responsive to high-frequency random fluctuations in excitation, simply by including PSC dynamics with a relatively long time constant of 20 ms. Depending on the frequency content of signals in a given circuit, this particular filtering mechanism might not be useful. However, there are other more sophisticated neural circuits that can perform, for example, band-pass filtering with any choice of corner frequencies [368]. This reinforces the conclusion that precise, reproducible behaviour can in theory arise from highly variable neural activity.

5.3.8 A Continuum with Rate Coding

The sections above consider the computations that can be performed on the basis of irregular firing patterns. A separate but related question is how such irregular firing patterns might arise from input signals that do not contain fluctuations at the same frequencies. Such a relationship would arise if a neuron's net driving current were an irregular function in the space of represented information. In this case, its firing pattern would be a deterministic function of the input to the ensemble, but it would appear irregular, even if the input changed smoothly over time.

This possibility is illustrated in Figure 5.9, in a model composed of leaky-integrate-and-fire (LIF) neurons. In this model, irregular patterns of somatic driving current are modelled abstractly. In a more detailed model, they might arise from weighted synaptic input (Chapter 3), possibly in combination with cell-intrinsic nonlinearities (Chapter 4). Analogously to previous sections, neurons that are post-synaptic to an ensemble of such irregularly-firing neurons can reliably extract represented signals, in the absence of both spike time coincidence and firing rate variations (Figure 5.9C).

Figure 5.9D shows a firing-rate histogram for one presynaptic model neuron, over thirty simulations in which the represented variable x had exactly the same trajectory. Despite this consistency in the input, there was no systematic variation in the firing rate over time, because the other represented variable y changed from trial to trial. This simulation illustrates that if it is not possible to control all of the variables to which a neuron is sensitive, then even in the absence of noise, relationships between a neuron's firing and the variables of interest can be completely obscured.

The pattern of driving current in Figure 5.9A is dominated by high-frequency fluctuations with respect to the represented variables (x and y). In contrast, the driving current of a 2-dimensional cosine-tuned neuron would be an inclined plane. Intermediate patterns are also possible. For example, a driving-current function might be inclined along the y-axis, but dominated by high frequencies along the x-axis. In general, a network could exhibit a continuum between timing and rate codes, through irregular current functions that are variously inclined along preferred directions, with variously-scaled high-frequency peaks.

5.4 Discussion

This chapter has shown that even in the absence of coarse rate variations, irregular firing patterns can drive nearly any given pattern of activity in a post-synaptic neuron. Importantly, such transformations can be obtained through learning. These results have two main implications for the interpretation of experimental data. First, a neuron's pattern of firing around an event may not have an obvious temporal relationship with the neuron's role in the event. For example, although a

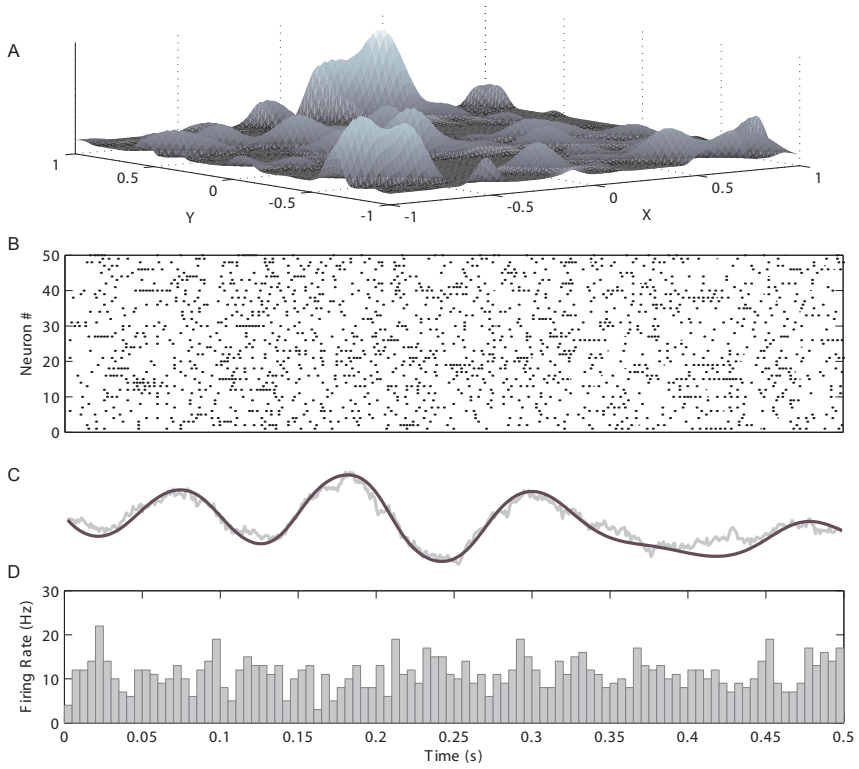

Figure 5.9: Temporal coding and decoding with LIF neurons. A, Net somatic current (arbitrary units) experienced by an example LIF neuron, as a function of two inputs (x and y). B, Irregular firing in 50 different neurons (each with different current functions) as inputs x and y vary at low frequency. C, Estimate of x decoded from activity of an ensemble of 1000 LIF neurons firing as in B. Black line indicates ideal decoding (post-synaptic current dynamics applied to input x). Gray line indicates the estimate of x by a neuron post-synaptic to the ensemble. This estimate is a weighted sum of post-synaptic currents generated by the firing of the ensemble. D, Firing-rate histogram showing a lack of mean firing rate dependence of an example neuron on input x, over 30 trials. In each trial x is identical, but y varies randomly.

group of neurons fires faster only at the end of a movement, subtle differences in spike timing between neurons may drive some aspect of movement initiation. This is particularly true with respect to irregular and highly stereotyped firing patterns, such as those arising in middle temporal responses to some visual stimuli [33] or in songbird vocalization [147] (although the same cannot be said if responses lack diversity across the population, e.g., see [311]). Furthermore, accuracy degrades gracefully with firing pattern variability, so that even patterns that are difficult to distinguish from random firing can drive relatively stereotyped activity. Therefore, the second main conclusion to be drawn from this chapter is that neither precise spike timing nor observable rate fluctuations can be relied on to expose all the significance of a cell's activity.

Although the focus of this chapter has been on projections from a single neural ensemble to a single post-synaptic neuron, the results also have implications for larger circuits. A single ensemble of neurons can drive different post-synaptic neurons in entirely different patterns (e.g., Figure 5.1C). Several hundred neurons driven in diverse patterns would form a rich basis from which to drive activity in a subsequent layer. Therefore, although it remains to study how errors propagate through multiple layers, the present results clearly apply to larger circuits as well as to single projections.

These findings are in general agreement with the results of Gütig and Sompolinsky [142] on the classification of firing patterns. If a neuron can be trained to spike in response only to selected population-temporal input patterns, as they have shown, then it would be expected that the same neuron could be made to exhibit arbitrarily chosen firing patterns by training it to respond only to selected short segments of a longer presynaptic pattern.

Medina et al. [251] present a model of a specific neural circuit that they take to function in similar manner to the abstract circuits in the present study. Theirs is a classical conditioning model, in which cerebellar granular cells respond to a conditioned stimulus with diverse temporal firing patterns. An unconditioned stimulus serves as a training signal, decreasing or increasing the strength of granular cell synapses onto Purkinje cells, depending on whether granular cell activity is coincident with the unconditioned stimulus or not. After training, Purkinje cells in effect decode a temporal prediction of the unconditioned stimulus from diverse granule cell firing patterns. Synaptic weights are modulated on the basis of a target output rather than error, so learning ends when some physiological parameter is saturated, rather than when error is minimized. Otherwise, this learning mechanism is analogous to the one presented here.

The present study is also conceptually related to the liquid-state machine [232]. The liquid-state machine relies on a diversity of neural responses to input, within a recurrent circuit, in order to approximate a broad class of temporal functions of the input. In contrast to the liquid-state machine (the neurons of which fire at fluctuating rates), the present study explores how computations are effected by firing statistics in the absence of large-scale rate fluctuations. This focus leads to

new implications (as describe above) with respect to the interpretation of electrophysiological data.

5.4.1 Effects of Firing Statistics on Performance

The relationships between the statistics of presynaptic firing patterns and the accuracy of PSC generation are remarkable in several respects. First, the irregularity of experimentally observed spike trains can provide a substantial functional advantage in terms of 1) the accuracy with which neurons can drive current in a postsynaptic cell and 2) the robustness of the current pattern to noise. For slowly varying current patterns, this advantage is even more pronounced with bursting neurons, highlighting a possible dimension in the functional relevance of burst firing that has received little attention (e.g., in [80, 94, 229, 310, 179, 196]).

Second, although greater numbers of neurons can drive current more accurately, very large numbers of neurons are not needed, even in the absence of precise spike timing or rate variations. As shown in Figure 5.3, 1500 irregularly and inconsistently firing neurons can drive useful PSC patterns. The degree of convergence onto most neurons is far greater than this. For example, some α-motoneurons receive about 50 000 synaptic inputs, and cerebellar Purkinje cells receive as many as 200 000. This indicates that multiple firing rate-independent signals could converge on a single neuron pool. Furthermore, because the same population firing pattern can induce vastly different currents in different cells (e.g., Figure 5.1C), the same small group of neurons could drive a wide variety of activity elsewhere, limited only by the number of different cells to which it projects.

Third, under the conditions studied here, errors in PSC are greater when the timing of presynaptic spikes is correlated. However, the increase in error is moderate when spike times are correlated at high frequencies. It is interesting to consider this result in relation to oscillations in local field potential (LFP), particularly in the context of motor control. Lower-frequency alpha and beta oscillations in motor cortical LFP usually disappear during movement and are sometimes replaced, around movement onset, by higher-frequency gamma oscillations [233]. Similar changes in LFP oscillations during movement occur in the cerebellum [297] and basal ganglia [67, 222, 79, 199]. Thus, patterns of LFP oscillation in motor areas during movement and rest coincide with patterns of synchrony that allow and preclude (respectively) the type of coding presented here, pointing to the possibility of a role for this type of coding in motor control.

Fourth, and finally, errors in pattern generation were dominated by high-frequency fluctuations, a point that is also relevant to motor control. For example, 75% of the error in Figure 5.2A was at frequencies above 100 Hz, much higher than the frequency content of skeletal movement. The frequency spectrum of the error is relevant in the context of motor control, because the relationship between myoelectric activity and muscle forces resembles a low-pass filter [285], and limb inertia has a further damping effect. Thus, most of the error observed in this study (i.e., error at

high frequencies) would not necessarily interfere with movement kinetics if it were present in a motor circuit.

5.4.2 Timing versus Rate

Each of the synthetic firing patterns used in this study was generated from a constant ISI distribution and in this sense has a constant mean firing rate. However, instantaneous rates fluctuated because the patterns (with the exception of those in Figure 5.2B) contained a range of ISIs. So, if these firing patterns were repeated over multiple trials, rate fluctuations would appear in the multi-trial spike histogram (although such fluctuations might be quite subtle, as in Figures 5.3 and 5.8). However, repeated task behavior does not guarantee that related neurons exhibit repeated patterns. This was illustrated in Figure 5.9, in the context of neurons with irregular tuning curves, but it is also true in other cases. For example, a neuron's activity may reflect something that varies from trial to trial, such as an error signal. Also, a neuron's firing pattern might contain information about a repeated feature of an event only when considered in conjunction with the firing patterns of other neurons [327]. Because instantaneous rate does not uniquely determine multi-trial rate, even if the neuron is noise free, and because it is otherwise indistinguishable from timing, this chapter uses the term "rate" only to indicate the inverse of the mean of the ISI distribution.

5.4.3 Limitations and Future Work

The most important limitation of this study is that the dendritic model used here assumes linear combination of currents, as might occur, for example, with synapses on separate distal dendrites [301]. Dendrites can also combine synaptic input in much more complex and varied ways, although some complexities of dendritic processing (including dendritic spiking) serve partly to compensate for passive cable properties rather than to implement nonlinear computations, as discussed in Chapter 3 [235, 236, 385, 318]. As noted in the Methods, the present results are relevant to any case in which PSC is a monotonic function of total conductance. For any target current, in such cases, there is a corresponding sum of conductances that will produce it. In more complex cases, the present results may only apply under limited conditions, for example, to activity within a single dendritic branch or within a certain voltage range. It is beyond the present scope to explore how these results interact with more detailed models of specific cell types, but it would be reasonable to expect that in many cases, sophisticated dendritic processing would enable further computations on the results of the computations modeled here. For example, several temporal current patterns that are generated by near-linear synaptic integration might converge to be combined multiplicatively. The possibility of such additional dendritic processing does not seem to affect the basic conclusion that arbitrary timing-based information can be exploited in a flexible manner, under very general circumstances.

One aspect of dendritic processing that would be particularly interesting to study, in relation to the present results, is variability in the dendritic membrane time constant (e.g., through neuro-modulation). Changes in membrane time constant would alter the temporal relationships between somatic currents arising from different parts of the dendritic tree. If weights were tuned in relation to one time constant, such changes would be expected to result in additional noise in the somatic current, at frequencies of about 50 Hz and higher. However, it might be possible to tune synaptic weights in order to exploit such changes functionally. For example, modulation of the time constant might synchronize or desynchronize distal excitatory inputs from more proximal inhibitory inputs, dramatically influencing the spiking pattern.

This chapter has shown that in principle, timing patterns can be exploited by the brain even if they are difficult to detect experimentally. This result is in a sense its own limitation, because it would be difficult to confirm that this was actually happening in a given circuit. A prerequisite would be that some functionality of a circuit could not be accounted for by firing rates or precise timing. Specific results of this chapter (e.g., relationships between error and firing statistics) may also help to resolve whether such a mechanism is feasible given other knowledge of the circuit. However, the only obvious way to test for this phenomenon directly is to perform large numbers of trials.

Another limitation of this study is that although a learning rule has been identified, which makes use of information that could plausibly be available at each synapse (i.e., each synapse does not need information from other synapses), this rule is speculative rather than being based on a known biological mechanism. It remains either to map this learning rule onto a demonstrated mechanism or to explore the viability of other rules, for example, rules based on rewards rather than error signals.

While the focus here has been on how an ensemble of neurons can produce a single pattern of PSC in a given cell, it is unlikely that a cell is dedicated to producing a particular pattern. Further work is therefore needed to explore how the present results generalize to the production of different current patterns in the same cell over short timescales, that is, without substantial changes in synaptic weights. There are several possibilities. For example, an ensemble could produce a family of pattern primitives, which could be separately gated to produce a wide range of PSC patterns. A circuit of this form might function as a repository of arbitrarily complex motor programs, with parameters varied through gating.

It may also be fruitful to further explore how the firing patterns that arise from varying input to a network could drive a useful set of outputs. Certainly, the firing patterns that are produced by two different inputs could produce essentially any two patterns of PSC. This is clear if one imagines that the spike pattern from 0 to 500 ms in Figure 5.1 is produced by one input, and the pattern from 500 to 1000 ms is produced by a second input. With a single set of synaptic weights, these two inputs result in two different current patterns. This remains true for

more than two inputs, but error rises roughly linearly with the summed duration of the input/output mappings. However, if firing patterns reflected only a few milliseconds' input, then multi input-multi output mapping might result in good piecewise approximations of a large family of desired outputs. This possibility is related to the liquid-state machine [232], but differs in a significant respect. Specifically, although computations in a liquid-state machine require traces of long-past inputs, the present suggestion is that a similar architecture without such traces may enable population coding of time-varying inputs without time-varying firing rates.

5.4.4 Population-Temporal Coding

The present results make it clear that patterns of irregular spiking, perhaps generated by recurrent circuit dynamics, can drive a wide range of time-varying activity in other cells. In this light, it may be reasonable to view any circuit that produces a temporal firing pattern, regardless of whether the pattern contains variations in firing rates, as being analogous to a central pattern generator. That is, such a circuit is a versatile, intrinsic source of time-varying activity patterns (although mechanisms of pattern modulation may be different from those of classical central pattern generators).

However, the ability of neurons to exploit timing-based information may have much broader uses. One interesting possibility is that a given pattern of input to a neuron might be analogous to the neuron's preferred direction, in a multi-dimensional population code. For example, suppose a neuron were to receive input from a number of synfire chains [103, 172]. The phase relationships among $n+1$ chains would span an n-dimensional vector space. Every vector in this space, that is, every possible list of phases, would correspond to a certain pattern of input to the receiving neuron. As the present results demonstrate, almost any such input pattern could be transformed into almost any pattern of current. Moreover, deviations from this input pattern, either in terms of phase relationships or spike-timing precision, would result in noisier current, much like deviations from preferred direction in a rate-based population code result in reduced current. An ensemble of neurons with different preferred phase relationships could support a population code over the space of phase relationships. The present results also suggest that a population code of this form could drive either a similar code in a receiving ensemble of neurons, or a rate-based population code (the latter is evident from the square-pulse example of Figure 5.1, in that a postsynaptic neuron would fire faster during the excitatory pulse). Further work is needed to verify that such a population code can be supported by realistic neuron models and to explore its computational power.

In conclusion, the results of this study suggest that neurons can use information contained in the timing of incoming spikes, under very general conditions. Synchrony is not needed, and specialized synapses, neurons, and circuit structures are

also unnecessary. Furthermore, incoming patterns can consist mostly of noise and can therefore be very hard to detect experimentally, yet still produce behaviorally useful patterns. Finally, timing-based information can be transformed into a wide variety of outputs, in a manner that seems to accommodate a versatile population code.

5.5 Appendix: Details of Power Analyses

The effect sizes for power analyses were derived from the smallest increases in the firing rates of a noisy excitatory population that could be expected to produce a spike in a cell post-synaptic to this population. For simplicity, it was assumed that PSCs would decay such that the post-synaptic cell would fire if it received more than a fixed number of spikes from excitatory sources within a 5-ms time bin. The rates of extra spikes and missing spikes in the postsynaptic cell were assumed to be the same, so that noise could be expressed as a single index, corresponding to the rate of mistimed spikes. For each spike in a postsynaptic cell, let n be the number of excitatory neurons converging onto the postsynaptic cell that have a slightly elevated, noisy rate increase that contributes probabilistically to the spike. The mean number of spikes in each bin, across these neurons, will be different for each trial. For large n, these trial means cluster around grand means in a Gaussian distribution with variance k/n (where k is the Poisson spike rate per bin). Reliability of postsynaptic spiking in this scenario will increase with greater differences between the grand means of the normal and elevated rates of presynaptic spiking. The grand-mean elevated rate of presynaptic spiking was set such that trial means for each bin crossed an intermediate threshold at a rate corresponding to a predetermined rate of mis-timed post-synaptic spikes. Because rates were elevated only in very short (5 ms) bins, this rate modulation can also be viewed as a noisy manipulation of spike timing.

These analyses result in estimations of the numbers of trials in various conditions, which provide a 0.8 probability of finding minimal rate elevations (if they exist), with a 1-way fixed-effects analysis of variance (ANOVA). The baseline and elevated rates were similar, so (because variance equals mean in a Poisson process) the ANOVA assumption of uniform variances was approximately satisfied. However, because the ANOVA relies on the sampling distribution of variances, which is sensitive to deviations from normality in the underlying distributions, results are presented from numerical experiments rather than from theoretical distributions. Each reported data point corresponds to the number of trials in each of a set of 1000 experiments, in which the null hypothesis (i.e., the hypothesis that there was no difference in firing rates across bins) was rejected between 799 and 801 times ($\alpha = 0.05$). The validity of the ANOVA with Poisson-distributed data in these circumstances was also confirmed, in that the null hypothesis was rejected at the $\alpha = 0.05$ level in roughly 50 of 1000 experiments in which there were no systematic rate differences, regardless of the number of trials in each experiment.

Chapter 6

Computation with Inhibitory Projections

In contrast with the cortex, the projection neurons of most basal ganglia nuclei are inhibitory. This property is central to conceptual models of basal ganglia function. For example, the classic Albin/DeLong model describes basal ganglia function in terms of a balance between inhibition and disinhibition of the output nuclei. In more recent computational models, it is generally assumed that if a group of neurons represents an item of information, the effect of inhibition is to reduce the represented value. It was shown recently [295] that excitatory projection neurons can subserve much more complex and varied computations (e.g. nonlinear and non-monotonic functions of the input), closely approximating any computation that can be achieved through a mix of excitatory and inhibitory synapses. This chapter shows that the same is true of inhibitory projection neurons, given certain other conditions that are also met in the basal ganglia. This observation suggests that the basal ganglia may be capable of much richer computations than previously recognized.

6.1 Introduction

From a theoretical perspective, the computational power of a neuronal projection is greatly enhanced if each of the presynaptic neurons can act of a source of both excitatory and inhibitory synapses. In particular, for any map subserved by n synapses with weights of uniform sign, there are 2^n maps with synaptic weights of the same absolute values but mixed signs. But weights of mixed sign are not physiologically realistic. Normally, a single neuron contains a single primary neurotransmitter (i.e. a single neurotransmitter that affects the trans-membrane currents of its targets), which has either an excitatory or inhibitory effect.[1] This appears to constrain the computational power of biological neural networks.

[1] There are exceptions. These include excitatory/inhibitory cotransmission in the retina [399] and possibly in the mammalian uterus [62]; the capacity for GABA to depolarize a cell, depending on resting membrane potential and internal Cl^- concentration [376, 69]; and receptor-dependent

However, it was shown recently [295] that any idealized projection model that contains synaptic weights of mixed sign can be transformed into a physiologically-realistic projection that is functionally equivalent. This transformed projection is consistent with typical cortical anatomy [346], in that 1) the primary projection neurons are excitatory, 2) these neurons synapse onto both excitatory neurons and inhibitory interneurons in the target area, 3) the inhibitory interneurons in turn synapse onto local excitatory neurons, 4) there is substantial convergence and divergence in the synapses between each group of neurons, and 5) about 20% of the neurons are inhibitory.

Interestingly, descending projections onto thalamic nuclei have a similar form [188].[2] Projections from the cortex onto the striatum are also similar, in that excitatory projection neurons synapse onto the striatal projection neurons, which in turn inhibit other projection neurons collaterally, and also onto fast-spiking inhibitory interneurons that inhibit medium spiny neurons in a feedforward manner [357, 299]. The striatal projection neurons are inhibitory, but this does not affect the function of the cortico-striatal projection.

In contrast with the typical cortical structure, the projection neurons in most basal ganglia nuclei are inhibitory. However, this chapter shows that because in most cases the target neurons are also inhibitory and tonically active [352], and have local collaterals, this connectivity is also functionally consistent with mixed excitatory and inhibitory synaptic weights. The projection from globus pallidus externus to the subthalamic nucleus is an exception, in that the target neurons are excitatory, but in this case parallel projections through additional globus pallidus neurons could theoretically have a similar effect. In the projection from striatum to SNc, the target neurons are dopaminergic. However, striatal neurons also synapse onto a smaller population of inhibitory interneurons, which in turn synapse onto the dopamine neurons [145].

These results imply that basal ganglia projections may underlie more sophisticated computations than previously recognized, particularly by the Albin/DeLong model and the more recent models that elaborate it.

6.2 Feedforward Excitatory Projections

This section reviews the transformation of a mixed-weight projection model to a more realistic model with excitatory projection neurons. The material in this section was introduced by Parisien et al. [295], elaborating on suggestions by Eliasmith & Anderson [111].

mixed effects of glutamate [193]. Some synapses can also switch rapidly between excitatory and inhibitory transmission, under control of brain-derived neurotrophic factor [391].

[2] Ascending projections usually have a different structure, in which dendrite-to-dendrite inhibitory complexes associate with individual excitatory synapses.

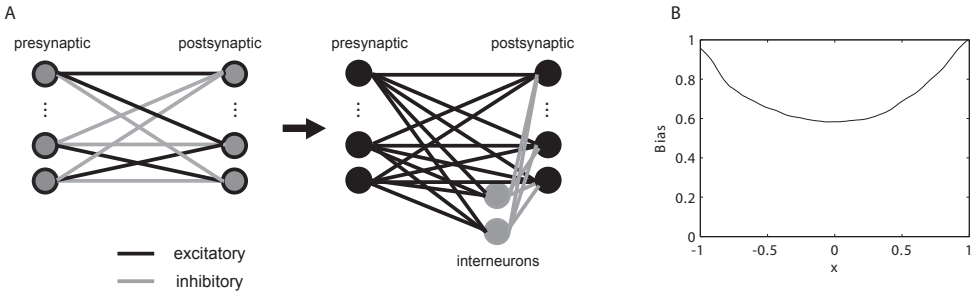

Figure 6.1: The excitatory Parisien transform. A, An idealized projection in which each pre-synaptic neuron can act as a source of both excitation and inhibition can is mapped to a physiologically realistic projection that performs the same computation. In the transformed projection, excitatory presynaptic neurons synapse both directly onto post-synaptic targets, and also indirectly through a small population of inhibitory interneurons. B, Shifting the synaptic weights in the main projection (so that they are all excitatory) introduces an excitatory bias current into the post-synaptic neurons. This bias current is a function of the variable $\mathbf{x}$ that is represented by the presynaptic ensemble. This same bias function is projected to the interneurons, which then offset the excitatory current by inducing an approximately-equal inhibitory current in the post-synaptic neurons.

Beginning with an idealized projection, in which presynaptic neurons make both excitatory and inhibitory synapses, Parisien et al. [295] define a transform to a realistic model in which each neuron is either excitatory or inhibitory. The transform consists of two steps. The first is to offset all of the original (mixed-sign) synaptic weights so that they become excitatory. This eliminates the mixed weights, but introduces extraneous excitatory current into the post-synaptic neurons. The second step is to cancel out this extraneous excitatory current by introducing inhibitory neurons. This could be done by introducing one inhibitory interneuron for each (now-excitatory) projection neuron. However, this would require as many inhibitory as excitatory neurons, contrary to cortical anatomy. In the method of Parisien et al., the inhibitory neurons instead encode all of the necessary bias as a population. This scheme is illustrated in Figure 6.1.

After the transformation, the synaptic weight between the i^{th} presynaptic neuron and the j^{th} postsynaptic neuron is

$$w_{ji} = w^o_{ji} + w^b_{ji},$$

where w^o_{ji} is the original (mixed-sign) synaptic weight, and w^b_{ji} is the positive bias that is added to this weight in order to make it excitatory. The only trick in defining the bias weight is that it must allow compensation by a correlated ensemble of interneurons. This correlation is what decouples the number of interneurons from the number of projection neurons. Parisien et al. achieve this by defining the bias

weight in terms of scalar encoders and decoders, as

$$w^b_{ji} = \tilde{\phi}^b_j \phi^b_i,$$

where (using the notation of the NEF, introduced in Chapter 3) $\tilde{\phi}^b_j$ is the bias encoder of the j^{th} post-synaptic neuron, and ϕ^b_i is the bias decoder of the i^{th} presynaptic neuron. The values of the bias decoders are not critical, although large differences between their magnitudes turn out to cause problems, so they are chosen to be uniform, i.e. $\phi^b_i = \phi^b$. These ϕ^b can be viewed as decoding a "bias function" $f^b(\mathbf{x})$ from the presynaptic neurons. With uniform ϕ^b, the form of this bias function is determined by the presynaptic neurons' tuning curves. For example, for cosine-tuned LIF neurons with diverse thresholds, this bias function resembles a parabola that is lowest around zero and highest at the extremes of the represented range (Figure 6.1B). The uniform ϕ^b are scaled so that this bias function has a maximum of one.

The bias encoder $\tilde{\phi}^b_j$ of the j^{th} post-synaptic neuron is then chosen to be as small as possible, such that $w_{ij} \geq 0$ for all i. This is achieved when

$$\tilde{\phi}^b_j = max_i \left(\frac{-w^o_{ji}}{\phi^b} \right).$$

Eliminating negative synaptic weights in this manner introduces an additional excitatory bias current $\tilde{\phi}^b_j f^b(\mathbf{x})$ into each post-synaptic neuron. To recover the transform $f^o(\mathbf{x})$ associated with the original mixed-sign weights w^o_{ji}, the presynaptic ensemble projects the bias function $f^b(\mathbf{x})$ to an ensemble of inhibitory interneurons. The interneurons have uniform encoders $\tilde{\phi}_k = 1$, and decoders that optimally approximate $-f^b(\mathbf{x})$, within the constraint that these decoders must all be negative. The interneurons then project this approximation $-\hat{f}^b(\mathbf{x})$ to the post-synaptic neurons, which scale it with the bias encoders $\tilde{\phi}^b_j$. Each post-synaptic neuron therefore receives the following map of the presynaptic represented variable:

$$f(\mathbf{x}) = f^o(\mathbf{x}) + \tilde{\phi}^b_j f^b(\mathbf{x}) - \tilde{\phi}^b_j \hat{f}^b(\mathbf{x}) \approx f^o(\mathbf{x}).$$

In other words, the shift in the synaptic weights of the main projection adds excitatory bias current, and the interneurons add approximately equal inhibitory current, so that the elaborated projection model has effectively the same synaptic weights as the original idealized projection.

As discussed in the introduction, the general structure of the resulting projection (e.g. excitatory neurons projecting onto both excitatory neurons and a smaller number of locally-connected inhibitory neurons, etc.) is very common.

6.3 Feedforward Inhibitory Projections

This section shows that the method described above extends in a straightforward manner to the case of inhibitory projection neurons. This case is less common

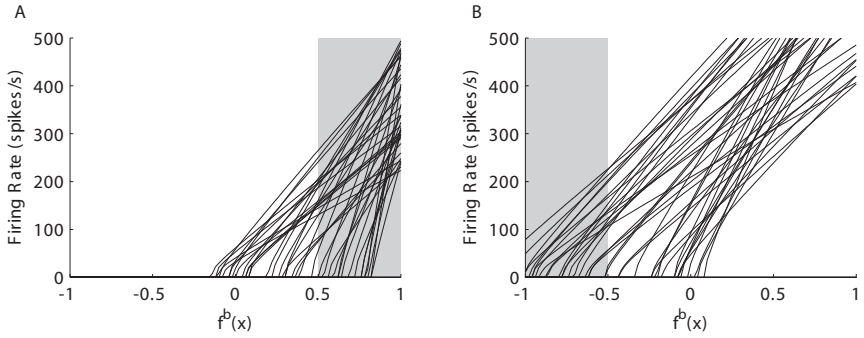

Figure 6.2: Interneuron tuning in the excitatory and inhibitory transforms. The shaded area indicates the normal operating range, and the lines show tuning curves of example neurons from interneuron ensembles. A, In the excitatory transform, the excitatory presynaptic neurons *increase* interneuron firing from *low* intrinsic rates (i.e. at $f^b = 0$). B, In the inhibitory transform, the inhibitory presynaptic neurons *reduce* firing activity from *high* intrinsic rates.

generally, but it dominates the basal ganglia – projection neurons of the striatum, globus pallidus, and substantia nigra compacta are all inhibitory.[3]

To transform an idealized (mixed-sign) projection into an inhibitory one, a *negative* bias is added to each of the original synaptic weights. In this case the bias decoders are uniform and negative, resulting in a bias function that is negative for all $\mathbf{x}$. The equation for the bias encoders actually remains the same, despite the fact that the largest-amplitude positive weight must be corrected in this case (rather than the largest-amplitude negative weight), because the bias decoders are negative. So again,

$$\tilde{\phi}^b_j = max_i \left(\frac{-w^o_{ji}}{\phi^b} \right).$$

The weight bias in this case introduces extraneous inhibitory currents. The bias function is also projected to tonically-active inhibitory interneurons, which fire more slowly as a result, and inhibit the post-synaptic neurons less. This balances the greater direct inhibition from the presynaptic neurons.

The tonic activity of the interneurons is critical, because reduction in this activity is needed in order to disinhibit the post-synaptic neurons (see Figure 6.2). The post-synaptic neurons must also be tonically active. Specifically, the tonic input from the inhibitory interneurons must be offset either by intrinsic currents or separate excitation.

Figure 6.3 compares an idealized mixed-sign projection with its transformation into both excitatory and inhibitory projections. In this example, the projection

[3]Incidentally, Purkinje cells, the projection neurons of the cerebellar cortex, are also inhibitory.

calculates a nonlinear and non-monotonic function of the presynaptic represented variable, illustrating that this type of computational flexibility is retained in both the excitatory and inhibitory cases.

The fact that the Parisien transform can be extended to inhibitory projections is not at all surprising. However, it is noteworthy, in that it calls into question one of the most common assumptions about how the basal ganglia work, i.e. that inhibitory projections correspond to a simple suppression of activity in the projection neurons of the target nuclei. This assumption is ubiquitous in action-selection models. For example, in one of the more sophisticated recent models [59], the influence of the striatum on the GPi (within the k^{th} action channel) is $-.54B_k^{SD}$, where B_k^{SD} is striatal activity. The present results suggest that essentially any $f(B_k^{SD})$ would be a potential alternative.

This is the main point of the chapter. The remainder of the chapter addresses relatively subtle questions of stability limits and performance.

6.4 Recurrent Projections

In a Parisien projection, the interneuron currents are slightly lagged in time behind the direct bias currents, because of the extra synapse in the pathway through the interneurons. This lag introduces an error, the magnitude of which varies with df^b/dt. In a feedforward network this error tends to be small, in part because excitatory synapses onto inhibitory neurons tend to have fast dynamics. However, as Parisien et al. [295] pointed out, the associated delay raises the possibility of instability in a recurrent network. They investigated this possibility using an integrator network as an example, and did not discover a stability problem. The integrator example is a reasonable choice, because by definition it operates on the border of instability. However, it remains possible that instability might arise in other types of recurrent networks.

This issue is particularly relevant to the inhibitory transform. In a recurrent inhibitory network (for example composed of laterally-connected neurons in the globus pallidus) it would be reasonable to expect that any threat to stability might be more pronounced. This is because the interneurons in such a network are of the same type as the post-synaptic neurons, so that all the synapses have the same dynamics, and the lag through the dis-inhibitory channel is relatively greater compared to that in the excitatory transform.

This section reconsiders the stability issue in light of this difference. It is shown that 1) the Parisien transform can lead to instability, even if the corresponding idealized circuit is stable, and 2) when the PSC time constant of the interneurons is as large as the others, the stability limits are narrowed. The consequence is that an all-inhibitory recurrent network with uniform PSC time constants can exhibit a restricted range of dynamics.

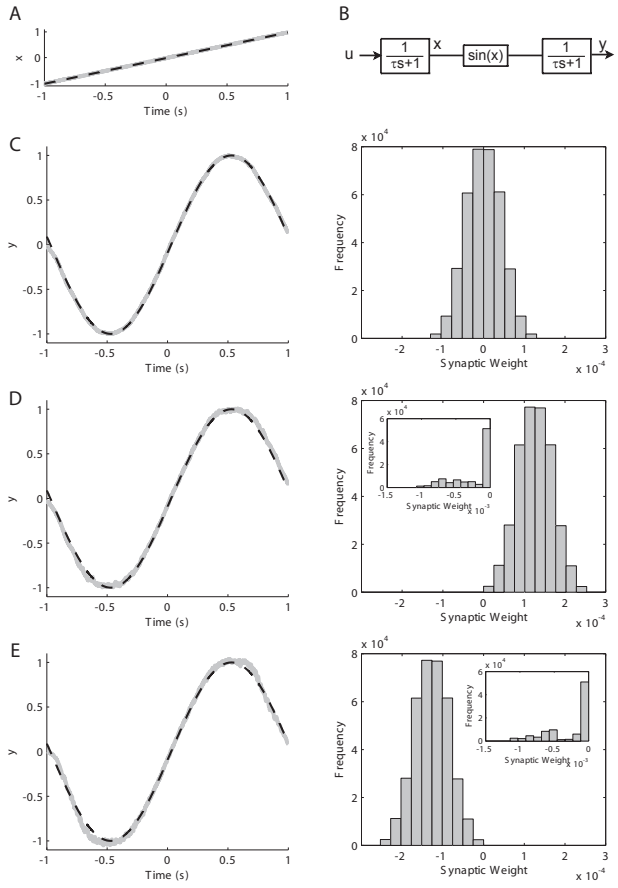

Figure 6.3: Example simulations illustrating that both the excitatory and inhibitory transforms can calculate non-monotonic functions. In each of the left panels, the black dashed line indicates the ideal value of a represented variable, and the gray line indicates its estimate, decoded from spiking activity in the corresponding ensemble. Each of these simulations was performed with ensembles of 600 presynaptic neurons, 600 post-synaptic neurons, and (in the transformed projections) 150 interneurons. A, The presynaptic ensemble represents an input variable that increases linearly with time. B, Diagram of the network structure, consisting of a single projection from a one-dimensional ensemble to another, in which the synaptic weights approximate the map $y = sin(x)$. C, Optimal linear decoding of y from spiking neural activity in the post-synaptic ensemble, with an idealized mixed-weight projection (MSE=.00018). The right panel shows a histogram of the synaptic weights in this projection. D, Optimal linear decoding of y from activity in the post-synaptic ensemble, after the excitatory Parisien transform (MSE=.0011). The right panel shows the shifted distribution of synaptic weights in the main projection (all above zero). The inset shows the distribution of synaptic weights in the projection from the inhibitory neurons to the post-synaptic neurons. E, As (D), but with the inhibitory Parisien transform (MSE=.0010).

6.4.1 New Instability Modes

In a recurrent network, the pre-synaptic and post-synaptic ensembles are the same, and will be referred to in this section as the primary ensemble (as opposed to the interneuron ensemble).

Ideally, bias in the direct feedback projection is cancelled out by feedback through the interneurons. However, the bias and interneuron feedback may be imbalanced due to distortion error in the interneuron ensemble, and also (when the represented value is changing) due to the additional lag in the path through the interneurons. This difference Δ^{di} between direct and indirect bias is the key to understanding how the network can become unstable. Both the direct and indirect bias affect the primary neurons through synapses. So the effect of this difference on the primary neurons at any given instant in time is $d = h(t) * \Delta^{di}$, where $h(t)$ is the post-synaptic current kernel, and $*$ denotes convolution. In other words, the firing of the primary neurons depends in part on the difference between direct and indirect bias, filtered by the post-synaptic current dynamics.

Recall that the bias encoders $\tilde{\phi}_j^b$ all have the same sign, so the firing rates of all neurons in the primary ensemble increase with increasing d. The bias function f^b was described above as a function of $\mathbf{x}$. However, it is really just a sum of the activities of neurons in the primary ensemble. So, because these neurons are also affected by d, the bias is more accurately described as a function of both $\mathbf{x}$ and d. Consequently, d can be viewed as a state variable that forms part of an additional feedback loop through the network, as illustrated in Figure 6.4. Accounting for this new state variable d, the system has dynamics

$$\begin{aligned}
\tau\dot{\mathbf{x}} =& A'\hat{\mathbf{x}}(\mathbf{x}, d) - \mathbf{x}, \\
\tau\dot{d} =& f^b(\mathbf{x}, d) - \hat{x}_2(x_2) - d, \\
\tau_2\dot{x}_2 =& f^b(\mathbf{x}, d) - x_2,
\end{aligned}$$

where x_2 is the variable represented by the interneuron ensemble, $\hat{x}_2(x_2)$ is the decoded estimate of x_2 from interneuron activity, and similarly $\hat{\mathbf{x}}(\mathbf{x}, d)$ is the decoded estimate of $\mathbf{x}$ from primary ensemble activity.

The stability problems are not obscured if the above model is simplified by assuming that $\hat{\mathbf{x}}(\mathbf{x}, d) = \mathbf{x}$. This assumption is reasonable in that 1) there are many primary neurons, with effectively-unconstrained synaptic weights, so the code for $\mathbf{x}$ is expected to be relatively accurate, and 2) moderate changes in d have very little effect on $\mathbf{x}$. This is because for any change in d, neurons with opposite preferred directions either increase or decrease their firing rates together. Thus some neurons code $\mathbf{x} + \Delta$ and others code $\mathbf{x} - \Delta$, and the net decoded value remains close to $\mathbf{x}$. This simplified system can be linearized around a nominal solution, as

$$\begin{aligned}
(\tau s + 1)\delta d =& \frac{\partial f^b}{\partial d}\delta d - \frac{\partial \hat{x}_2}{\partial d}\delta x_2 + \frac{\partial f^b}{\partial \mathbf{x}}\delta\mathbf{x}, \\
(\tau_2 s + 1)\delta x_2 =& \frac{\partial f^b}{\delta d}\delta d + \frac{\partial f^b}{\partial \mathbf{x}}\delta\mathbf{x},
\end{aligned}$$

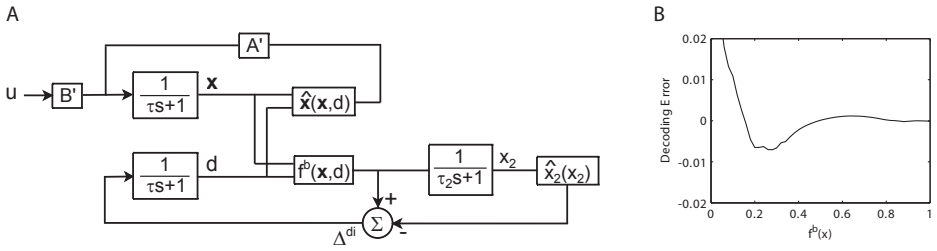

Figure 6.4: Sources of instability. A, A diagram showing the standard NEF feedback network (using post-synaptic current dynamics as memory, as discussed in Chapter 3), and elaborated to include the additional factors introduced by the Parisien transform. In particular, neuron activity in the primary ensemble gives rise to the bias function f^b, which feeds back both directly *via* the main projection, and also indirectly through the interneurons. The indirect route introduces a lag, and an additional decoding error. As in Chapter 3, PSCs are modeled with first-order exponential dynamics (in the diagram, PSC dynamics correspond to blocks with transfer functions that contain the Laplace variable s in the denominator). Note that the two dynamic blocks $1/(\tau s + 1)$ both correspond to synapses onto the primary neurons. These blocks are separated according to the logical distinction between the $\mathbf{x}$ and d state variables. Physically, the direct feedback (which includes both $\hat{\mathbf{x}}$ and f^b) corresponds to both of the feedback paths in the diagram that do not pass through x_2. B, Decoding error in an example interneuron ensemble (150 neurons). The constraint on the sign of the decoders makes the decoding error relatively large, particularly near zero.

where $\delta d = d - d_0$ and $\delta x_2 = x_2 - x_{20}$ are small deviations around the corresponding states of the nominal system. To simplify the notation, let $\alpha = \partial f^b / \partial d$, and let $\beta = \partial \hat{x}_2 / \partial x_2 - 1$ (note that β is the slope of the decoding error function shown in Figure 6.4B). The linearized system has the feedback matrix

$$A = \begin{bmatrix} \frac{(\alpha-1)}{\tau} & \frac{-(1+\beta)}{\tau} \\ \frac{\alpha}{\tau_2} & \frac{-1}{\tau_2} \end{bmatrix}.$$

The eigenvalues of the system are $\lambda = [(\frac{\alpha-1}{\tau} - \frac{1}{\tau_2}) \pm \sqrt{(\frac{\alpha-1}{\tau} - \frac{1}{\tau_2})^2 - 4(\frac{\alpha\beta+1}{\tau\tau_2})}]/2$. An unstable eigenvalue ($\lambda > 0$) will exist if either $\frac{1}{\tau_2} < \frac{\alpha-1}{\tau}$, or the square root is real and $\beta < -1/\alpha$. As anticipated, the former case corresponds to interneuron PSC dynamics that are too slow, relative to PSC dynamics in the primary ensemble. To return to the main issue at hand, i.e. the effect of uniform PSC dynamics among the interneurons and post-synaptic neurons, the relatively larger τ_2 relative to τ in this case means that correspondingly smaller α can be tolerated. In the latter case, a sufficient negative slope in the decoding error of the interneuron ensemble causes a self-perpetuating divergence between the direct and indirect feedback. Ultimately all the neurons in the network saturate at their maximum firing rates.

The magnitude of α is a critical parameter. It varies with x_0 and d_0, but unfortunately its range is hard to define. This is because it is a function of the bias encoders, which depend on the synaptic weights, which in turn depend in complex ways on the tuning curves of the primary ensemble. However, α generally increases (endangering stability) with increases in the absolute values of the entries in A'. Thus, counter-intuitively, an idealized circuit with large negative eigenvalues will be unstable in Parisien form. There is no simple expression for this stability boundary, but all else being equal, it is inversely proportional to τ_2/τ. Examples of systems on either side of a stability boundary are given in Section 6.5.2, below.

6.5 Optimization

This section describes several modifications that improve the performance of the Parisien projection. Performance optimizations may seem tedious beside the main theoretical results, but they are important for understanding practical constraints on circuit function, and for understanding whether a hypothesized function is viable. While the changes described below lead to substantial improvements, further improvements are undoubtledly possible, and suggestions are given for future investigations along these lines. For simplicity, the discussion is in terms of the excitatory transform, which is more intuitive, but the same methods apply to the inhibitory transform as well.[4]

[4]These optimizations have been implemented for both excitatory and inhibitory transforms in the open-source simulation package Nengo (www.nengo.ca), the initial release of which was developed in parallel with this thesis.

6.5.1 Minimizing Interneuron Error

Ignoring dynamics for the moment, the main error introduced by the Parisien tranform is distortion in the interneuron representation. This error tends to be large relative to that of the presynaptic ensemble, because 1) the interneuron ensemble contains fewer neurons, and 2) its decoders are sign-constrained, so that decoding is very probably less accurate than the unconstrained optimum.

A preliminary improvement can be made by parameterizing the interneuron tuning curves so that decoding error is low over a broad range of input. When the interneurons have fairly linear response functions above threshold, and thresholds uniformly distributed from just below zero to just below one, the distortion error is relatively low from about 0.2 to 1 (see Figure 6.4B). Probably a further improvement could be achieved through gradient descent on the neuron parameters, but this possibility is not explored here.

A second way to reduce the decoding error is to normalize the bias function, to match the low-error region of the interneurons. Parisien et al. [295] set the magnitude of the uniform bias decoders so that the bias function peaks at one. With diverse cosine-tuned LIF neurons, the bias function then typically ranges from about 0.5 to 1 (see Figure 6.1B). It is advantageous not only to avoid high-error regions in the interneuron representation, but also to span the entire low-error region. This is helpful because if the bias output must be multiplied by >1 to achieve this, then the interneuron output must be divided by >1, along with the associated error. If f^b_{min} and f^b_{max} define the range of the bias function, and i_{min} and i_{max} define the range of the well-coded region of the interneuron ensemble, then the projection of the bias function to the interneuron ensemble must be multiplied by a factor a, and biased by a factor b, so that $af^b_{min} + b = i_{min}$ and $af^b_{max} + b = i_{max}$. If $(f^b_{max} - f^b_{min}) < (i_{max} - i_{min})$, as is typically the case, then $a > 1$. But in any case a is as large as possible. To compensate for these changes, the post-synaptic neurons then require intrinsic bias current $-b/a$, and the output of the interneuron ensemble must be scaled by $1/a$. Again, this means the distortion error introduced by the interneurons is divided by as large a value as possible.

A final and more subtle improvement can be made by minimizing $f^b_{max} - f^b_{min}$. This can be accomplished by replacing the uniform bias decoders with the non-uniform decoders. The new decoders must be non-negative, and must minimize this range without increasing the bias encoders of the post-synaptic ensemble. Figure 6.5 shows an optimized bias function that was found using constrained gradient descent ($f^b_{max} - f^b_{min}$ is reduced from 0.422 to 0.067). Figure 6.5 also reproduces the Parisien-transformed sine function decoding of Figure 6.3, with and without these optimizations. The optimizations decrease the mean-squared error from $1.1\text{x}10^{-3}$ to $2.4\text{x}10^{-4}$, which approaches the error of $1.8\text{x}10^{-4}$ in the idealized mixed-weight projection.

The changes described above reduce the error that is introduced by the interneurons. It is possible that further improvements could be achieved if the synaptic

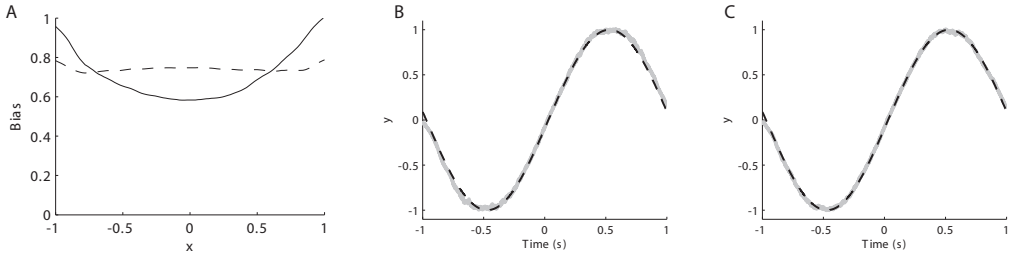

Figure 6.5: Optimization of the Parisien transform. A, Bias function with uniform decoders (solid) and with decoders optimized for flatness, as described in the text (dashed). B, Decoding of post-synaptic activity in the excitatory transform example from Figure 6.3 (non-optimized). MSE=1.1x10^{-3}. C, As (B), but with optimization of the bias function, and of the scaling in the indirect projection. Performance in this case approaches that with the idealized mixed-weight projection (MSE=2.4x10^{-4}, as opposed to 1.8x10^{-4} in the idealized projection).

weights between the interneurons and the post-synaptic neurons were optimized to compensate for error in the main projection. However, errors in the main projection typically have high frequency, so this would probably require that the interneurons have irregular tuning curves.

6.5.2 Balancing Feedback

There is another degree of freedom that can strongly influence the performance of recurrent circuits. It is well known from linear systems theory that a given transfer function (which defines dynamic input-output behavior) can be realized by a variety of state-space models. The implication for neural networks is that if a circuit is hypothesized to have certain input-output dynamics, there are a variety of feedback structures that could lead to these dynamics. It turns out that some of these structures may be unstable after a Parisien transform, while others perform well.

The key point from linear systems theory is that if one begins with a state-space model,

$$\begin{aligned}\dot{\mathbf{x}} =& A\mathbf{x} + B\mathbf{u},\\ \mathbf{y} =& C\mathbf{x} + D\mathbf{u},\end{aligned}$$

then a change of basis of the state variables does not affect the input-output behavior. A change of basis can be effected by any invertible map P, so that the state vector on the new basis is $\mathbf{x} = P\mathbf{x}'$. This results in the new state equations

$$\begin{aligned}\dot{\mathbf{x}}' =& P^{-1}AP\mathbf{x}' + P^{-1}B\mathbf{u},\\ \mathbf{y} =& CP\mathbf{x}' + D\mathbf{u}.\end{aligned}$$

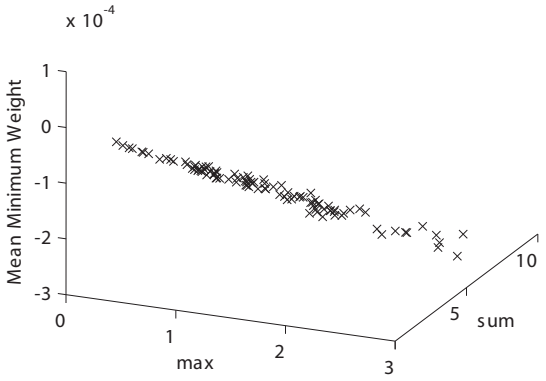

Figure 6.6: The key stability parameter α (see text, section 6.4.1) is correlated with both the maximum of the entries in $|A'|$ and their sum. This plot is based on a projection from one two-dimensional ensemble to another, which calculates a linear map with the matrix A'. The plot shows means (across post-synaptic neurons) of the minimum weights of all synapses onto a neuron in an idealized projection. These values determine the bias encoders, which determine α (which is itself not plotted because it varies over $\mathbf{x}$ and d). Each 'x' corresponds to a different A' with random entries.

The NEF realization of this state model will be farthest from the feedback instability described in Section 6.4 when the feedback matrix $A' = \tau P^{-1}AP + I$ minimizes α, which is correlated with both the sum and the maximum of the entries in $|A'|$ (Figure 6.6).

As an example, Figure 6.7 shows simulations of two different Parisien-transformed NEF implementations of a band-pass Butterworth filter, with transfer function $H(s) = \omega^2 s/(s^2 + \sqrt{2}\omega s + \omega^2)$, where ω is the corner frequency. These two implementations correspond to different canonical realizations of the transfer function, specifically the controller-canonical realization and the modal-canonical realization (see [70]). The modal-canonical realization is stable. However, despite the fact that the eigenvalues are identical in the controller-canonical realization, both the sum and the maximum of $|A'|$ are higher, and the Parisien transform of this network is unstable. Considering only the second realization in this case would have led to the incorrect conclusion that the neurons could not realize this transfer function.

Perhaps one could define a neural-canonical realization that minimizes error within stability constraints. The effects of errors from other sources within the network, particularly the distortion and noise arising from the main projections, are also influenced by the choice of state-variable basis. In order to find the optimal realization given a certain number of neurons, the sum and maximum of $|A'|$ would have to be considered simultaneously with the numbers of neurons coding each state variable, and the matching of state variable ranges with accurately-represented

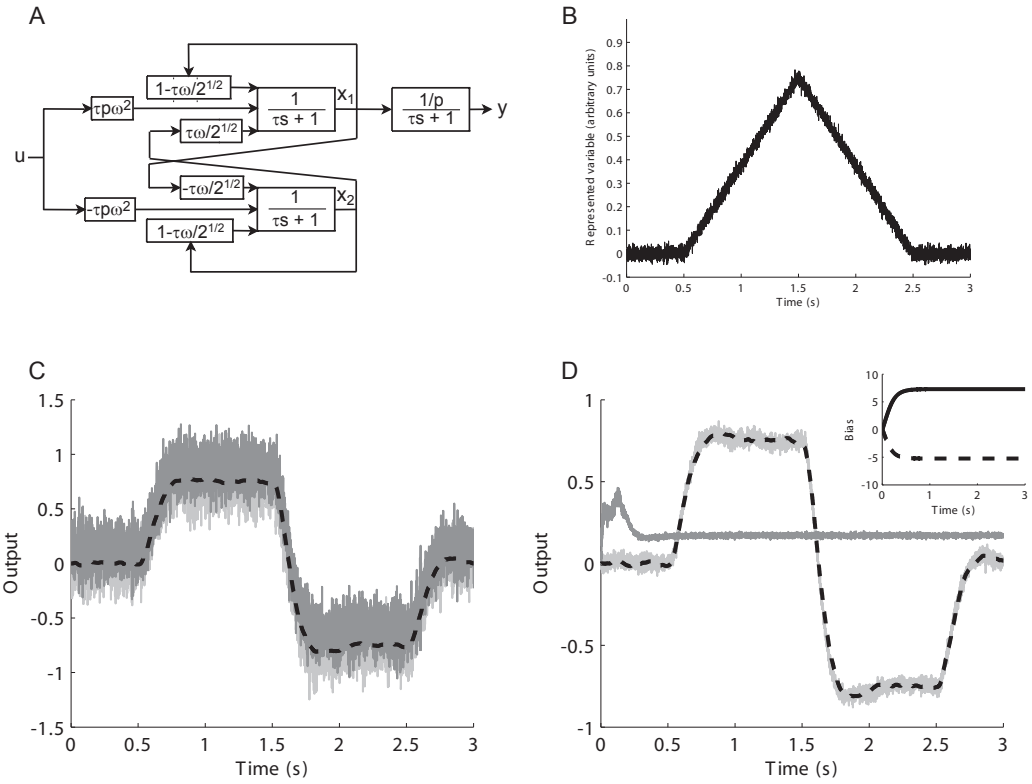

Figure 6.7: Instability in one of two Parisien-transformed feedback networks that have the same nominal eigenvalues. In their idealized form, both of the networks implement the same band-pass Butterworth filter, and both are stable. A, Diagram illustrating the modal-canonical realization of the filter, as a recurrently-connected ensemble that represents the state vector $\begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$ (ω is the corner frequency, and p is a parameter that scales the state variables so that they remain within the represented ranges). The controller-canonical realization is analogous, but with different feedback terms. B, A noisy ramp signal used as input to each network. C, Decoded output of the modal realization. The dashed line indicates ideal output, the light gray line indicates decoded output of the idealized mixed-weight model, and the overlapping dark-gray line indicates decoded output of the Parisien-transformed model. D, As (C) but with the controller-canonical realization. The performance of the idealized controller-canonical network is superior to that of the modal-canonical realization. However, the Parisien-transformed network is unstable. The inset illustrates the mode of instability, which is due to the slope of the interneuron decoding error. The indirect bias output by the interneurons (dashed) is smaller than that of the direct bias (solid), and this difference feeds back on itself until all of the neurons in the network saturate at high rates.

ranges of the ensembles. This is an interesting but non-trivial problem that must be deferred for future work.

6.6 Discussion

The main point of this chapter is that the inhibitory projections of the basal ganglia can, in theory, support a great variety of computations. One consequence is that relatively-accurate decoding of represented variables is possible, compared with the more constrained decoding implied by simple inhibition. Another consequence is that these projections can support a broad range of nonlinear and even nonmonotonic functions of represented variables.

As one example, suppose a part of the striatum were to contain two competing channels that represent the selection of different actions. A nonlinear map in the two-dimensional space of these actions could trivially implement a winner-take-all mechanism in a feedforward manner. Specifically, the striato-pallidal projection could achieve this through a Parisien transform, by decoding the function $[\ x_1 > x_2 \quad x_2 > x_1\ ]^T$, where x_1 and x_2 correspond to the degree of selection of the two actions within the striatum. In contrast, specific striatal tuning would be required in order to approximate this type of function with sign-constrained decoders.

It was noted above that the Parisien transform is not affected by the synaptic action of the *post*-synaptic neurons, i.e. whether they are excitatory, inhibitory, or modulatory. This means that the excitatory transform describes the cortico-striatal projection, despite the fact that the post-synaptic neurons (the medium spiny neurons) are inhibitory. For the same reason, the inhibitory transform describes the pallido-thalamic projection – in which most pallidal neurons terminate both directly onto thalamic projections neurons and indirectly through local circuit interneurons [174] – despite the fact that the projection neurons of the thalamus are excitatory.

The inhibitory Parisien transform results in a network in which all neurons are inhibitory. This raises the possibility that a recurrently-connected inhibitory population (e.g. within the globus pallidus) might project arbitrary functions onto itself, and thereby form a dynamical system that approximates any set of ordinary differential equations (as described by the NEF; see Chapter 3). This is an interesting possibility, but the results of Section 4 suggest that such a network would have a relatively narrow region of stability, limited by the strong feedback that is needed to achieve network dynamics that are faster than PSC dynamics.

Finally, even in static conditions, the performance of a Parisien-transformed network is somewhat impaired relative to an idealized mixed-weight model, due to distortion error introduced by the interneurons. The degree of impairment is an important theoretical consideration with the potential to constrain hypotheses about network function. However, it varies with several factors that can be seen as

degrees of freedom in the transform. Section 6.5 introduced several modifications that minimize impairment relative to the ideal case.

6.6.1 Challenges for Experimental Validation

If we observe anatomy of essentially the right form, how can we tell whether a Parisien projection is present? Are there any experimental observations that would rule it out? This type of validation presents a difficult problem, because the transform is robust to a variety of changes that result in different predictions about connectivity and firing patterns.

In a specific Parisien-transformed model (e.g. Figure 6.7), the firing patterns of the interneurons and post-synaptic neurons are different. This seems to suggest that one could develop a Parisien-transformed model of a specific system, and then check experimentally whether a minority of neurons exhibit firing patterns that look like the interneuron firing patterns. The first problem arises in defining the size of this minority. Parisien et al. [295] assume 20% interneurons (to match the proportion of inhibitory neurons in the cortex), but the proportion is less critical for performance than the absolute number. Several projections might share the same interneuron ensemble. With more projections sharing the same interneurons, the performance would degrade gradually, because finer differences in the value of the bias function would become significant. Eventually the physical limit of convergence onto the interneurons would be reached, but this limit could be surpassed if the bias function were coded by only a subset of the correlated presynaptic neurons. In summary, the proportion of interneurons required for the Parisien transform is not well defined.

A second issue is that the distinct firing pattern of the interneurons is not well defined. Parisien et al. [295] assume for convenience that the bias decoders are uniform, but this assumption is not critical. Different bias decoders would result in a different bias function, and consequently a different pattern of interneuron activity.

Finally, in the inhibitory case, there is no reason the interneurons and post-synaptic neurons have to be distinct groups. A single recurrently-connected, multi-dimensional ensemble could operate in the same manner. This further confounds expectations about classes of firing patterns in the network.

6.6.2 Conclusion

This chapter has argued on theoretical grounds that the inhibitory projections of the basal ganglia may be capable of sophisticated computations, much like the excitatory projections of the cortex. Ironically, in terms of information processing, this makes one of the most striking physiological features of the basal ganglia appear almost inconsequential. These theoretical results hint at a vast array of

unexplored possibilities in basal ganglia function. Further exploration of these possibilities should be grounded in experimental validation of the basic theoretical results. However, as outlined above, experimental validation will not be straightforward.

Chapter 7

Cell-Intrinsic Firing Dynamics

Large network models usually treat neural activity as a static function of synaptic input. However, the majority of real neurons exhibit dynamic responses to input, such as adaptation or bursting. These firing-rate dynamics are difficult to reconcile with a straightforward view of population coding, in which neurons encode the information they receive as a list of firing rates. This chapter discusses four different ways in which spiking dynamics can be reconciled with population coding, in various circumstances. First, in some cases spike-rate adaptation can interact with recurrent connections, to approximate any network dynamics that can be described by a set of explicit ordinary differential equations. Secondly, diverse firing dynamics across an ensemble can span a space of transfer functions, any of which can be realized by a different set of synaptic weights. Thirdly, it is shown that simple neuron models can decode not only the instantaneous input from a dynamic population response, but also the input history, and functions of the input history. Finally, in contrast with the other cases, spike dynamics are essentially irrelevant for neurons that use an averaging code. Rather than affecting the represented value in this case, they modulate the degree of noise in the representation over time. Together, these results point toward a richer and more realistic view of population coding, which embraces the prominent dynamic properties exhibited by many neurons.

7.1 Introduction

The majority of neurons respond dynamically to constant input. For example, pyramidal neurons (the main neurons of the cortex) respond to a step increase in driving current with initially-rapid firing, which then adapts gradually to a lower steady state [248]. A large minority of thalamic relay neurons operate in two modes, one of which is to produce a high-frequency burst in response to a stimulus [380]. In the cerebellum, both Purkinje cells and granule cells burst rhythmically with prolonged depolarization [230, 88]. Medium spiny neurons, the projection neurons of the striatum, exhibit hystersis [276], in that they self-stabilize through intrinsic currents into low and high membrane-potential states.

There are single-cell models that reproduce these dynamics well (e.g. [176, 275]), and some of them have been incorporated into larger network models, resulting in complex and apparently-realistic network activity (e.g. [177, 180]). However, a limitation of these network models is that they do not perform any obvious computation.

Conversely, network models that do perform explicit computations usually ignore cell dynamics. This is true of models in the artificial neural network tradition (e.g. [157]), and of physiologically-inspired population-coding models (e.g. [286]). More surprisingly, models that focus specifically on the role of network dynamics in computation also usually ignore dynamics at the cell level. For example, many such models, following Abbott [1], approximate a neuron as a single-time-constant dynamic process (which approximates post-synaptic current dynamics), in series with a static spiking nonlinearity. Models of this form have been used to study plasticity, winner-take-all competition, and more recently the role of chaos in computation. Attractor dynamics constitute another computational mechanism in recurrent networks (e.g. as the substrate of content-addressable memory), and have been studied extensively [160, 110, 20], but again with little attention to the cell dynamics that undoubtedly influence them (but see [365]).

This is not to say that the role of spiking dynamics has been ignored. Thalamic bursting has been proposed to underlie rapid direction of attention [221], and intrinsic bursting in pyramidal neurons may contribute to large-scale oscillations [77]. In pyramidal cell models that can produce both spikes and bursts, bursts signal slightly different input events than single spikes, and have more reliable timing [196]. Furthermore, some forms of adaptation optimize information transmission in the face of changing input statistics [113].

In summary, cell-intrinsic spiking dynamics have well-understood causes [178], and a variety of implications for computation, but there is room for strengthening the links with population coding and computation in general. Furthermore, the fact that such a prominent feature of neuron behaviour is ignored in so many computational models casts some doubt on the validity of these models, and suggests that efforts to more thoroughly reconcile cell dynamics with other network properties might lead to new insights.

The Neural Engineering Framework (NEF; [111]) uniquely integrates network dynamics with population coding. As discussed in Chapter 3, this is achieved by treating represented variables as the state variables of linear systems theory, and replacing the integral in the standard state equations with a low-pass filter that models post-synaptic currents. As originally described, this approach shares with the work of Abbott [1] and others the assumption that neuron dynamics are dominated by post-synaptic current dynamics, i.e. that neurons will spike at a constant rate when they are driven by constant current. Notably, this assumption was subsequently relaxed in the working memory model of Singh & Eliasmith [341], which incorporates neurons with spike-rate adaptation. They found that adaptation, in conjunction with two-dimensional tuning curves, reproduced several types of dy-

namic responses observed in the prefrontal cortex during a working-memory task. However, while adaptation was important in this study for matching the model neurons' responses to electrophysiological data, memory traces were maintained by the model only after its neurons had adapted quite strongly, so that adaptation itself did not play an integral role in the network's computation.

This chapter considers the relationship between cell and network dynamics from a series of different perspectives. Adaptation dynamics are similar to post-synaptic current (PSC) dynamics, in that they have a low-pass component. So a sensible starting point for the present work is to adapt the NEF approach in such a way that network dynamics rely on adaptation as a form of memory. Section 7.2 shows that this substitution is possible in some cases, and furthermore that adaptation and PSC dynamics can also interact cooperatively. These results begin to systematically reconcile cell dynamics and computation. However, this approach is limited, in that it applies only to adaptation. Furthermore, it requires linear adaptation, at a uniform rate across an ensemble.

As a first step in moving beyond these limitations, Section 7.3 shows that a population with non-uniform linear firing-rate dynamics spans a space of transfer functions, any of which can be realized by a different set of feedforward synaptic weights. It is then shown that diverse ensemble dynamics allow decoding not only of the instantaneous input to an ensemble, but also of past inputs, and functions of past and present inputs. Several examples are discussed, including two that may play an important role in basal ganglia dynamics: the rebound burst of subthalamic nucleus neurons, and the hysteresis of medium spiny neurons of the striatum.

Finally, it is shown that in the averaging code (introduced in Chapter 4), variations in firing rate over time have little effect on the represented value, instead modulating the amplitude of the decoding error. In this context, an adapting neuron initially contributes to a very accurate code, which is then relaxed over time, sacrificing accuracy for energy efficiency.

7.2 Firing Dynamics can Provide Dynamical System Memory

This section considers firing-rate adaptation as a substrate of dynamical system memory. Recall from Chapter 3 that for any set of explicit ordinary differential equations, there is a family of neural circuits that has approximately the same dynamics, over some range of the state variables. The NEF shows how to substitute post-synaptic current dynamics in place of integration, in order to find this family of circuits. This section shows that a similar substitution can be made with firing-rate adaptation dynamics, provided 1) adaptation dynamics are linear, and 2) all neurons in the ensemble adapt with the same time constant.

The first step is to introduce a standard model of adaptation, and show that

it can meet these conditions, while retaining the tuning-curve diversity needed for linear decoding.

7.2.1 Uniform Linear Adaptation

The adapting leaky-integrate-and-fire (ALIF) model provides a good approximation of the adaptation effects observed in more detailed compartmental models [207, 275]. There are several variations of this model, but the following discussion is based on a version in which adaptation is driven by an unspecified chemical species N [64], the concentration of which varies as

$$d[\mathrm{N}]/dt = -[\mathrm{N}]/\tau_N + A_N \sum_k \delta(t - t_k).$$

Here [N] is the concentration of the chemical species responsible for adaptation, τ_N is a time constant of decay of [N], and $A_N > 0$ is an increment in [N] with each spike. This model can also be expressed in terms of firing rates (rather than spikes) as

$$d[\mathrm{N}]/dt = -[\mathrm{N}]/\tau_N + A_N r(u)$$

where $r(u)$ is the firing rate.

If these neurons are parameterized so that their unadapted (onset) firing rates vary nearly linearly with driving current, then each neuron behaves like a linear band-pass filter with a non-zero pole, i.e. output consists of low-pass plus band-pass components. If α is the derivative of the unadapted firing rate with respect to driving current, then as long as the neuron's firing rate does not drop to zero, it will have linear first-order dynamics with time constant

$$\tau_A = (1/\tau_N + \alpha g_N A_N)^{-1},$$

where the product $g_N[\mathrm{N}]$ is the conductance underlying adaptation (g_N is a coefficient that scales this conductance).

The diverse tuning curves that are required for linear function decoding result in diverse α across the ensemble. However, the other parameters (τ_N, g_N, and A_N) can co-vary with α, to yield uniform τ_A.[1] In this case the neurons have uniform, linear dynamics, and the ensemble as a whole has the same dynamics as each neuron.

These parameters also influence the steady-state adapted firing rate for a given input. Computations that exploit adaptation will be more accurate if the firing rate adapts more strongly (conversely, weak adaptation will be obscured by firing-rate noise). On the other hand, if adaptation is too strong, then a sudden drop in drive will cause the firing rate to drop to zero, which will change the time constant. For a given τ_A, the degree of adaptation is maximized within this constraint if 1) $\tau_N = \tau_A(b/c + 1)/2$, where b is the neuron's intrinsic bias current, and c is the net

[1] Non-uniform dynamics are discussed later, in Section 7.3.

synaptic current flowing into the neuron per unit of encoded scalar variable (over the range -1 to 1), and 2) $A_N = (1/\tau_A - 1/\tau_N)/\alpha$.

It will be shown shortly that the band-pass dynamics of adaptation can play a role that is analogous to that of PSC dynamics in a recurrent network. In the mean time, it is worth noting that short-term synaptic depression has similar band-pass dynamics, and could potentially play a similar role. However, in contrast with adaptation, the feedforward dynamics caused by synaptic depression are inherently nonlinear.

7.2.2 Synaptic Depression is Non-Linear

Synaptic depression is usually a pre-synaptic phenomenon that arises from depletion of the readily-releasable pool of synaptic vesicles. In simple models, the readily-releasable pool is depleted sharply with each spike, and replenished with exponential dynamics [405]. The dynamics of the readily-releasable pool are therefore closely analogous to those of [N] (above). If the synaptic weight with a full readily-releasible pool is w, then the effective synaptic weight at any instant is $wS(t)$, where $S(t)$ is the remaining proportion of the pool at time t. The state equations that relate the net input current u to weighted output y are then

$$\dot{S} = (1 - S)/\tau_S - FSr(u),$$

$$y = wSr(u),$$

where τ_S is the time constant with which the readily-releasable pool is replenished, F is the proportion (between 0 and 1) of the pool that is depleted with each spike, and $r(u)$ is the firing rate as a function of net input u.

Importantly, despite many similarities with adaptation dynamics, both the dynamic and output equations are non-linear functions of the firing rate. The dynamic nonlinearity is weak when F is low (although this also results in weaker depression). However, the output nonlinearity is unavoidable, and it implies that synaptic depression can only produce linear, band-pass behavior when the derivative of the firing rate is zero. Thus the intuitive notion that synaptic depression produces derivative-like output [4] is only very roughly accurate.

7.2.3 Adaptation Supports Integration

This section examines the viability of firing-rate adaptation as a general memory for network dynamics, using the example of integration. Integrator networks play important roles in working memory [261, 136] and oculomotor control [336]. More generally, integrators are of interest because they can serve as the foundation for other types of dynamics (see Chapter 3).

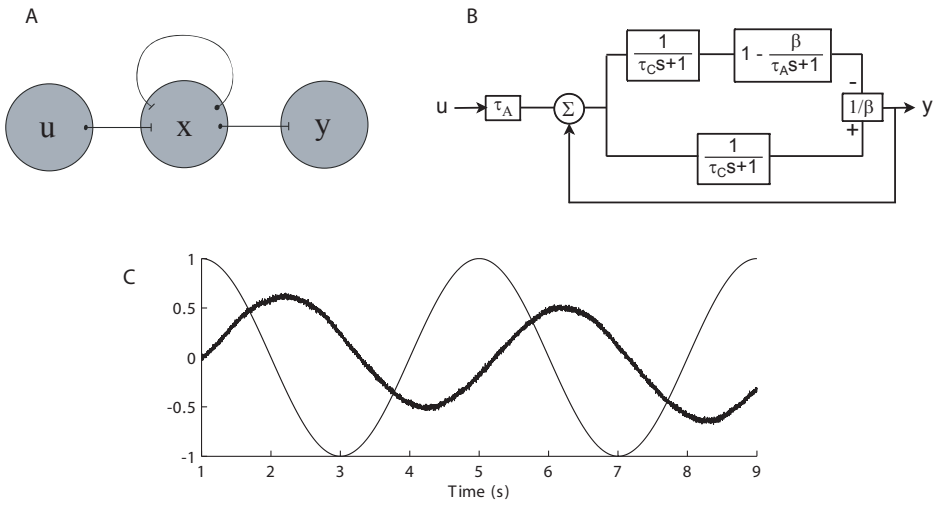

Figure 7.1: Integrator based on firing-rate adaptation. A, Sketch of the network structure, with each circle indicating an ensemble of neurons. An input ensemble (u) projects to the integrator ensemble (x). The integrator ensemble consists of both adapting and non-adapting neurons. B, Block diagram highlighting the dynamic elements of the network. The two parallel feedforward paths correspond to the adapting and non-adapting neurons. The transfer function $1/(\tau_C s + 1)$ models the dynamics of the post-synaptic currents in synapses onto the integrator neurons. The transfer function $1 + \beta/(\tau_A s + 1)$ models adaptation. C, Simulation with slow sinusoidal input. The simulated networks consists of 1500 adapting and 500 non-adapting neurons, with $\tau_A = 0.5s$ and $\tau_C = 0.005s$. The smooth line is the input, and the noisy line is decoded spiking activity of the output ensemble (y).

Figure 7.1 shows a network that integrates its inputs over time using firing-rate adaptation as memory. With uniform, linear adaptation dynamics (as described above), an adapting ensmble has the feedforward transfer function

$$H_A(s) = 1 - \frac{\beta}{\tau_A s + 1},$$

where β parameterizes the degree of adaptation of the represented variable. This band-pass response must be converted to a low-pass response in order to apply the NEF methods. A net low-pass response can be obtained if additional, non-adapting neurons lie in parallel with the adapting neurons (see Figure 7.1B). The full feedforward transfer function $H(s)$ also includes the PSC dynamics (which are assumed to be much faster than adaptation dynamics), i.e.

$$H(s) = \frac{1}{(\tau_C s + 1)(\tau_A s + 1)},$$

where τ_C is the time constant of the post-synaptic current.

The feedforward dynamics are now second-order, but this does not greatly effect the behavior of the integrator. If the neural input and feedback matrices are chosen

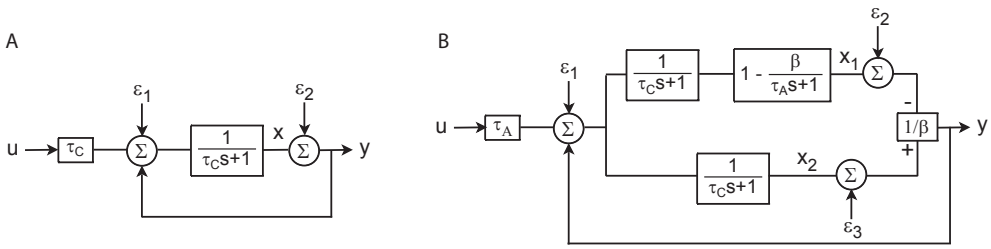

Figure 7.2: Error in PSC-based (A) *vs.* adaptation-based (B) integrators. The dynamic blocks correspond to PSC dynamics and adaptation dynamics (as in Figure 7.1), and ϵ_i are additive errors in the decoding of ensemble activity.

(using the methods in Chapter 3) as $A' = 1$ and $B' = \tau_A$, the dynamics can be described using two state variables, as

$$\dot{x}_1 = \frac{x_2 + \tau_A u}{\tau_C s + 1},$$

$$\dot{x}_2 = \frac{x_1}{\tau_A s + 1}.$$

The eigenvalues of these linear equations are $\lambda = \left\{0, -(\tau_A^{-1} + \tau_C^{-1})\right\}$. Thus one mode decays rapidly, and the other integrates. Figure 7.1C shows a simulation of this network, constructed from spiking ALIF neurons.

A question that immediately comes to mind is how well this integrator performs, compared to the PSC-based integrator. In the PSC-based integrator, the major source of error is distortion in the representation of the integrated value. This representational distortion can be modeled as additive error that is a function of the represented value, as shown in Figure 7.2A. The transfer function of this integrator, including distortion errors, is

$$y = \frac{u + \epsilon_1 + \epsilon_2/\tau_C}{s},$$

where ϵ_1 is the error in the decoding of input-ensemble activity, and ϵ_2 is the error in the decoding of integrator ensemble activity. The time constant of post-synaptic current decay is typically between 0.005s and 0.1s, much less than 1. Thus if ϵ_1 and ϵ_2 have similar magnitude, the error at the output is dominated by ϵ_2. Furthermore, if ϵ_2/τ has a magnitude greater than u, then the network ceases to integrate, and drifts toward an attractor in the error function [111].

The adaptation-based integrator has one potential advantage, in that adaptation dynamics can be much slower than post-synaptic current dynamics (on the order of seconds). In this case, ϵ_2 is divided by a much larger value. However, the adaptation-based circuit also has a number of disadvantages. Distortion error magnitude decreases with increasing population size n, as approximately $1/\sqrt{n}$ [111].

Given n neurons, an adaptation-based integrator must divide them between adapting and compensating cohorts, so that the distortion of each signal is higher. Furthermore, the outputs of the adapting and compensating populations are summed, and amplified by at least a factor of 3 (because firing rates cannot adapt by much more than one third, without the risk that they will drop to zero with a rapid drop in input; this means $\beta \leq \frac{1}{3}$ in the adaptation transfer function). Finally, the adaptation process itself might introduce further sources of error that are not present in non-adapting neurons.

Another difficulty is that the adapting and compensating populations must actually encode larger values than the population in the PSC integrator. The population in the PSC integrator must be able to represent the range of values corresponding to the integral (i.e. values of y). In the adaptation circuit, the populations must represent the value $\tau u + y$. Unfortunately, distortion error scales linearly with the range of values that an ensemble must encode. For large enough τ, this means that ϵ_2 will scale linearly with τ, exactly cancelling out the advantage of the longer time constant.

Despite these disadvantages, an adapting integrator might out-perform a PSC integrator in a circuit in which the integrals are large compared to the inputs, i.e. when the term $\tau u + y$ is dominated by y. Interestingly, this scenario does not correspond to the two most clearly-established examples of integrators in the brain, i.e. the oculomotor integrator (which integrates large saccade signals), and working memory, in which the integrals have the same magnitudes as the inputs.

In summary, firing-rate adaptation can serve as a substrate of integration. However, it probably does not confer a performance advantage over PSC-based integration, except in limited special cases.

7.2.4 Limitations of Adaptation-Based Memory

An ideal integrator can serve as the foundation for a variety of dynamical systems, through additional feedback. This is analogous to the realization of dynamical systems using low-pass synaptic dynamics (discussed in Chapter 3), only in this case there is no need to compensate for the filter (so $A' = A$ and $B' = B$). Unfortunately, the additional pole in the non-ideal adaptation-based integrator can cause instablility if the eigenvalues of the ideal system are large and negative. This phenomenon is analogous to the instability discussed previously (Chapter 6) in relation to feedback projections through interneurons, so it will not be discussed in detail.

The picture that emerges from the above analysis is that it is possible for adaptation dyamics to serve as the memory of a dynamic circuit, but quite specific circumstances are required. In particular, the analysis assumed linear adaptation dynamics that were uniform across an ensemble, and required that non-adapting neurons operate in parallel with the adapting neurons. Even under these strict conditions, the dynamics are at best slightly distorted compared with those of the PSC-based memory, and adaptation-based memory does not work at all for fast

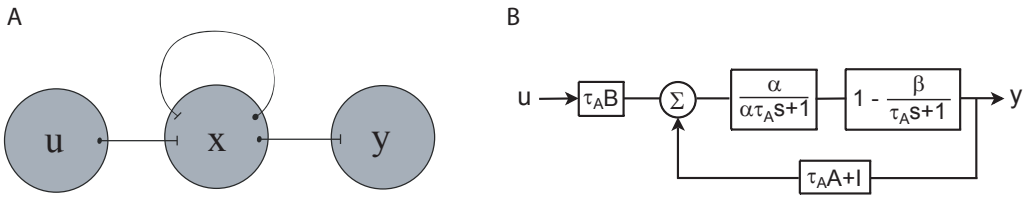

Figure 7.3: Stable integrator based on matched adaptation and PSC dynamics. A pole and zero cancel when $\alpha = 1/(1-\beta)$, leaving net first-order feedforward dynamics.

dynamics, due to additional stability limits. So while adaptation is a possible mechanism for memory in dynamic networks, it is not a particularly good one.

Still, many neurons adapt. If adaptation is not a good substrate for more general circuit dynamics, can it at least co-exist with them? Interestingly, there is network structure in which adaptation and PSC dynamics combine in such a way that the system as a whole does not have any extraneous modes, and therefore does not suffer stability problems.

This structure is shown in Figure 7.3. Here, ensemble adaptation dynamics are again modelled as $H_A(s) = 1 - \beta/(\tau_A s + 1)$ (where β is the degree to which the represented variable adapts), and the PSC dynamics are modelled as $H_C(s) = 1/(\tau_C s + 1)$, where the PSC time constant is expressed as a proportion of the adaptation time constant, i.e. $\tau_C = \alpha\tau_A$. This network depends on a specific relationship between the degree of adaptation, and the ratio between the two time constants. Specifically, it is necessary that $\tau_C = \tau_A/(1-\beta)$. (There is no trouble obtaining this relationship with realistic neuron parameters, because both τ_A and β depend on the neuron's unadapted tuning curve, which in turn depends on synaptic weights.) When this is the case, a pole and zero cancel, and the feedforward transfer function simplifies to $H(s) = 1/(\tau_A s + 1)$. The ensemble then has feedforward dynamics of the same form as the the PSC dynamics that served as memory in Chapter 3. So as in that case, the network can approximate any set of explicit ODEs, and (provided the pole and zero cancel exactly) there is no hidden unstable mode.

To summarize, this section has shown that adaptation can variously serve as a substrate for dynamical system memory, or contribute to this memory without corrupting the dynamics. However, in order to reach these conclusions, it was necessary to assume that the adapting neurons have linear unadapted response functions, and dynamics that are uniform across an ensemble. This does not rule out the possibility that nonlinear neurons could play similar roles, but it is not obvious how they could. The remaining sections propose alternative roles for cell-intrinsic dynamics that do not share these requirements.

7.3 Firing Dynamics can Span Transfer Functions

In the models of the previous section, all of the neurons adapted with the same time constant τ_A, so the represented variables also decayed with this time constant. However, there are often substantial differences between the dynamic responses of different neurons in a group. For example, neurons in the globus pallidus exhibit diverse dynamics, including both bursting and adaptation [203, 271], with varying parameters that may relate to variations in ion-channel density [135]. Subthalamic nucleus neurons also have diverse intrinsic dynamics, which are modulated differently in different cells by perfusion of dopamine-receptor agonists [44].

Gerstner & Kistler [132] point out that the transfer function of an ensemble of neurons with non-uniform linear dynamics can be estimated using system identification methods. This requires assumptions about synaptic weights and the distributions of dynamic properties. Of course, if the dynamics are diverse, then different assumptions about synaptic weights could result in identification of a different system. This is not a limitation of the method – it just means that an ensemble of neurons with diverse dynamics spans a space of transfer functions. A Monte Carlo estimate of this space could be obtained by performing system identification repeatedly, using randomized weights within plausible ranges. But in using this method with realistic numbers of neurons (hundreds or thousands), some interesting transfer functions might emerge very rarely, because of the large number of degrees of freedom.

Another approach is to hypothesize a transfer function, on the basis of the large-scale behaviour of the circuit, and test whether the neuronal population can realize it. This is analogous to the method for finding the optimal weights for approximating a static representation, except that the time dimension must also be considered. If the neurons were truly linear, this could be done using the neurons' impulse responses. However, even nominally-linear neurons saturate with large enough inputs, so step responses within the represented domain are more useful. The synaptic weights by which an ensemble optimally approximates hypothesized responses to steps (from 0 to various values of x) can be found by simulating the ensemble with step inputs, and minimizing the error,

$$E = \frac{1}{AT} \int_{\mathbf{x}} \int_{t} [\mathbf{x} f(t) - \sum_i a_i(x,t)\phi_i]^2 dt d\mathbf{x},$$

where $f(t)$ is the hypothesized unit step response, $a_i(x,t)$ are the neurons' responses, ϕ_i are the decoders, and A and T are the size of the represented domain, and the time span over which integration is performed, respectively.

Analogously to the static case discussed in Chapter 3, a concise description of the space of possible transfer functions can be obtained by principal components analysis. As an illustration, Figure 7.4 shows the first few principal components of an ensemble of adapting LIF neurons with diverse τ_A. The principal components span variations in both static computations and dynamics.

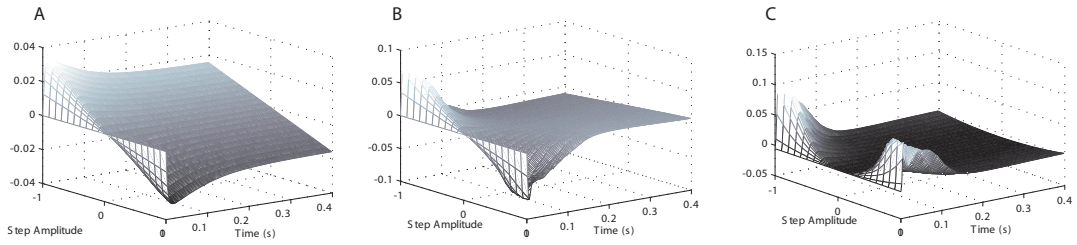

Figure 7.4: The first three principal components of the step response of an ensemble of near-linear adapting LIF neurons. In this ensemble there is variation in both adaptation dynamics and in the slopes of the onset response functions. Consequently, the principal components span both different dynamics (compare left and centre panels) and different static functions (compare centre and right panels).

If these neurons instead had uniform τ_A, and diverse degrees of adaptation, one of the principal components would be a pure low-pass response. If such an ensemble appeared in the adaptation-based integrator of the previous section, then decoding this response would eliminate the need for the parallel non-adapting neurons. In other words, varying degrees of adaptation are needed in order to decode a low-pass response, but these need not include zero.

7.3.1 Interaction between Forward Transfer Functions and Recurrence

If such an ensemble were to project recurrently onto itself, then the dynamics of the resulting network would depend on both the decoded transfer function and the feedback strength. For example, with feedback scaling a, and decoded ensemble transfer function $H(s)$, the transfer function of the recurrent network would be

$$\frac{x}{u} = \frac{1}{H^{-1}(s) - a}.$$

Importantly, higher-order dynamics in $H(s)$ do not confer the same level of generality to the feedback network as single-time-constant dynamics. For example, if $H(s) = 1/(s^2 + \sqrt{2}\omega s + \omega^2)$, i.e. a Butterworth filter with corner frequency ω, then

$$\frac{x}{u} = \frac{1}{s^2 + \sqrt{2}\omega s + \omega^2 - a}.$$

The feedback strength a influences the system poles, but the mean of the poles is $-\omega/\sqrt{2}$, regardless of a. So while diverse cell dynamics confer flexibility on the feedforward ensemble dynamics, exploiting this flexibility may constrain the ensemble's recurrent dynamics.

7.3.2 Non-Linear Firing Dynamics

The above approach only makes sense if the firing dynamics are linear, so that a response to any input within saturation limits also defines the cell's responses to other inputs. Linear dynamics just add one dimension to the neurons' tuning curves, i.e. the dynamic kernel. In contrast, if a cell has nonlinear dynamics, then its response to one input signal may indicate very little about its response to other input signals. So there is no simple kernel dimension that can be considered in order to account for the dynamics.

If the dynamics are only mildly non-linear (e.g. ALIF dynamics with saturating response functions), it is still possible to approximate a linear transfer function for some limited range of inputs (not shown). However, the approximation will be poor for novel inputs, i.e. for inputs outside the range of those for which optimal weights are obtained. Describing a nonlinear ensemble response with a linear transfer function may or may not be useful, but in any case optimal weights found using mathematically convenient inputs (e.g. steps) are not likely to perform well with natural inputs. Conversely, if optimal weights are found for realistic inputs, the resulting code will probably break down for step input, although this might be less of a modelling limitation than a limitation of real neural systems, many of which perform poorly when driven with un-natural stimuli [314].

If the firing dynamics are highly non-linear, this approach can still tell us about the range of ensemble responses that can arise over a very restricted range of inputs. This information may be useful if the neurons tend to receive stereotyped input signals (e.g. saccade-related bursts).

Example: Non-Linear Rebound Bursting in STN

Neurons in the subthalamic nucleus fire at a low intrinsic rate without synaptic drive. If these neurons are hyperpolarized from rest, they stop firing, and then fire a burst of action potentials after hyperpolarization is relieved. This burst is due to calcium channels that open when the membrane potential is low, and then close slowly (over a fraction of a second) when membrane potential rises again into the spiking range. These bursts endow the the firing response with a brief memory.

Figure 7.5 shows the principal components of responses of a population of subthalamic nucleus neurons to a specific input (a pulse), based on a single-compartment conductance model from Terman *et al.* [358]. Depending on synaptic weights, the influence of these responses on a post-synaptic neuron could be any linear combination of these principal components.

Limitations

If the neurons' dyamics are linear then the above approach, i.e. approximating hypothesized ensemble dynamics as a linear combination of neuron dynamics, provides

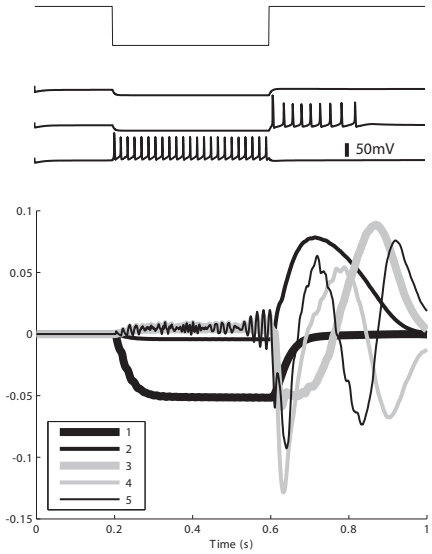

Figure 7.5: Rebound bursts of a population of subthalamic nucleus neurons following a pulse input (top). Different neurons in the population scale the input differently, leading to a variety of responses. The middle traces show membrane potential of three example cells. The bottom traces show the first five principal components of the population response, which include a variety of rebound responses on various time scales. This variety arises entirely from differences in input scaling, which determines the peak of the rebound-related calcium conductance in each neuron, and consequently the duration of the rebound burst.

very general information about the range of possible ensemble responses to input. However, if the neurons' dynamics are non-linear, this approach can only indicate the range of ensemble responses to a restricted set of inputs (e.g. saccade-related bursts, or other stereotyped signals).

7.4 Firing Dynamics can Encode History

Nonlinear cell dynamics are fairly common, so it would be surprising if their only effect was to corrupt ensemble dynamics. Furthermore, the previous sections have only considered linear *ensemble* dynamics, but nonlinear ensemble dynamics may also be important. Is there a better way of looking at nonlinearities?

In general, in order to account completely for nonlinear dynamics, it would be necessary to drive the neurons with all possible patterns of input. However, a specific given ensemble may have similar responses to many distinct input patterns. If a neuron's dynamics have a short memory, so that its activity only depends on input in the recent past, then the firing rate can be viewed as a function of input over some recent interval. Furthermore, if the neuron is not too sensitive to rapid fluctuations, then it will be possible to approximate its activity as a function of the input at a finite number of previous time steps. The neuron's tuning curve can then be reconceptualized as a function in this higher-dimensional space of past and present inputs. The tuning curve in this space accounts for the neuron's dynamics. Furthermore, as with static tuning curves, synaptic integration in post-synaptic neurons can be viewed as combining these tuning curves to approximate varied functions of the input history.

Of course, there is no guarantee in advance that this new way of looking at neuron responses will reveal a redundant population code. The input-history space may be higher-dimensional than the number of neurons in the ensemble, so that the neurons are largely independent. On the other hand the neurons' responses may be too highly correlated, so that different input histories can not be distinguished.

However, this new approach is appealing because 1) it is analogous to static computation *via* static tuning curves, and 2) it involves explicit consideration of the space of input features that determine nonlinear cell responses. These input features may be important determinants of network function, so it is worth exploring an approach that emphasizes them.

7.4.1 Example: Rebound Bursting Revisited

The rebound bursts of neurons in the subthalamic nucleus are considered again here, in order to introduce this approach. The low-threshold calcium channels responsible for bursting open and close with a time constant of about 50ms, so a neuron's response depends mainly on input within the last few hundred milliseconds.

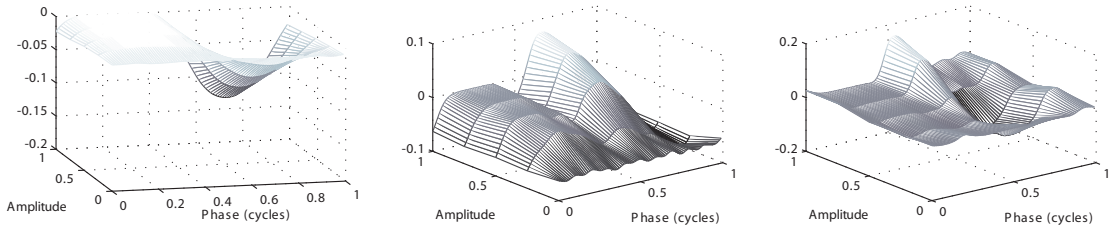

Figure 7.6: First three principal components of STN population response to 2Hz sinusoidal input. The population consists of twenty neurons, which differ only in terms of input scaling and bias. To obtain the principal components, responses of each neuron over three cycles were tiled, and spike outputs were filtered ($\tau_{PSC} = 10ms$). Each output was then further filtered with a Gaussian kernel (SD = 0.05 cycles), as an approximation of spike jitter in a larger ensemble of similar neurons. The first PC resembles a rectified version of the input. The higher peak in the second PC leads the input peak, due to rebound bursting. The peak in the third PC coincides with the strongest part of the rebound burst. The same ensemble was also simulated with rebound-burst conductances turned off. The first PC of this non-dynamic ensemble (not shown) was similar. However, the peak in the second PC did not lead the input, and there was no PC comparable to the third PC of the dynamic responses.

Intuitively, one might expect that the response could be well-approximated as a function of perhaps 5-10 historical dimensions. However, a restricted, two-dimensional input history will be considered, to allow visualization. Figure 7.6 shows the first three principal components of an ensemble response to 2Hz sinusoidal input. The history dimensions correspond to amplitude and phase of the input.

What dimension do these dynamics really have? This can be discovered by driving the neurons with higher-dimensional random input, and finding out how well tuning curves of various dimensions can fit the responses. This procedure (Figure 7.7) reveals that a neuron's response to sixty seconds of 0-5Hz band-limited input is well approximated by a four-dimensional response function.

7.4.2 Non-Linear Decoding of History

The principal components shown above indicate the functions of input history that can be decoded by linear synaptic integration. If physiological constraints on decoding are ignored, it is clear that nonlinear dynamics might convey very rich information about the input to an ideal observer. Non-linear synaptic integration allows a neuron to use this information in a more flexible manner, to extract an approximation of any function of input history, provided the ensemble response to

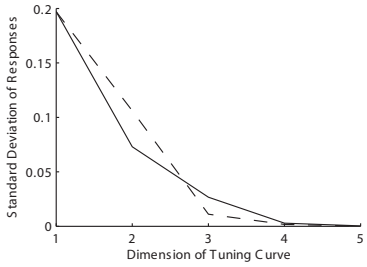

Figure 7.7: The response of a subthalamic nucleus neuron model [358] to band-limited drive (0-5Hz) is four-dimensional. Recent (250ms) input history during a long simulation was divided into 1-5 bins. The average input within each bin was taken as a different dimension of input history. In separate analyses, the bins had either equal time (dashed line), or time over which an exponential decay with $\tau = 50ms$ had an equal integral (solid line). The slow burst-related gating variable (the model's main dynamic element) at each instant was fit with a piecewise-constant function with eight steps in each history dimension. This function approximated the neurons' response function in the history space. The plot shows the standard deviation of the actual value of the gating variable around this n-dimensional function, as a proportion of its range over the whole simulation.

history is invertible (Chapter 4). If the ensemble response is not invertible, i.e. if two different patterns of input result in the same ensemble firing rates, then this approximation is restricted to functions that have the same value for these two input patterns.

Determining whether the ensemble response is invertible is not as straightforward as might be hoped. One complication is that the neuron responses are noisy, and since the tuning curve over finite-dimensional history is only an approximation, which may break down for novel input patterns, fluctuations around this curve may well be non-Gaussian, may have non-uniform variance, etc. If the distributions of possible ensemble responses at two nearby points in the history space overlap, this raises the possibility of a small error in the estimate of these points. On the other hand, if the distributions of two distant points overlap, a large error is possible. (Although low-pass post-synaptic current dynamics would tend to mitigate brief excursions in the estimate.) One way to avoid this type of problem (at least in a model) is to decode not a function of individual presynaptic activities, but of their principal components. Each of the large principal components is a sum of activities that is relatively robust to noise. If the principal components can be inverted, a function decoded from them should be similarly robust.

Figure 7.8 illustrates non-linear decoding of the history of input to a population of adapting leaky-integrate-and fire neurons with diverse adaptation time constants. The averages of input history in separate time bins (relative to the simulation time t) are decoded, as a systematic way of exploring decoding accuracy as a function

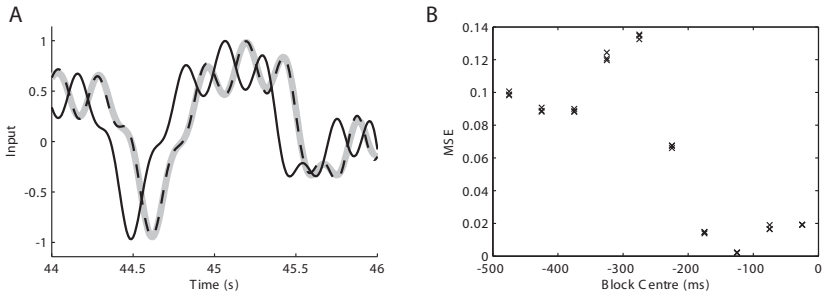

Figure 7.8: Non-linear decoding of input history from ALIF population response to band-limited white noise. The ensemble consists of 500 ALIF neurons with adaptation time constants that range from 0.02s to 0.2s. A sigmoidal backpropagation network with 50 hidden units was used as a model of a nonlinear dendritic tree (as described in Chapter 4). The first seven principal components of the ALIF responses were used as input to the network (rather than individual ALIF responses), in order to make the estimate more robust to noise. Forty seconds of simulation results were used as training data (sampled at 100Hz). The results presented are from ten seconds of separate testing data. A, Nonlinear decoding of the average input in a bin from $t - 150ms$ to $t - 100ms$. The solid black line shows a segment of input from the testing data set. The solid gray line shows the correct average over the history bin, and the black dashed line shows the decoded estimate of this history from the ALIF responses. B, Mean-squared error in decoding of 50ms bins, from $t - 500ms$ to t, with an input signal of mean-squared amplitude 1. Three separate models were trained for each bin, and the error of each estimate over the testing data is shown as an "x". The decoding is accurate over $200ms$ of history, which corresponds to the time constants of the slowest-adapting neurons.

of the lag time. The fact that history can be accurately decoded over a wide range of time lags suggests a great deal of flexibility in decoding functions of this history.

7.4.3 Infinite-Dimensional History

A neuron's dynamics can only be treated as a function of finite-dimensional input history if the neuron has a short memory, and a band-limited response to input fluctuations. While this is true for most neurons (e.g. adapting and bursting neurons), the medium spiny neurons of the striatum can exhibit hysteresis, depending on the concentration of extracellular dopamine. Over a wide range of net synaptic input, their firing rates may be higher or lower, depending on whether the level of input has crossed a high or low threshold more recently. Consequently, the activity of these neurons may depend on inputs as far back in time as dopamine concentration has been elevated (perhaps farther, due to residual decaying hysteresis with low

dopamine). So the dynamic response can not be well-described by a function in a finite-dimensional space. This leaves us without an obvious way to systematically discover which functions of input history can be decoded, but individual hypotheses can be examined case-by-case.

As an example, Figures 7.9 and 7.10 illustrate how an ensemble of hysteretic medium spiny neurons can code whether a sequence of events has occurred. Neurons in the model represent values in a two-dimensional space, which allows visualization of the tuning curves. Each neuron's synaptic drive is modelled abstractly as a two-dimensional Gaussian function in this space, and the neuron's activity is a nonlinear and dynamic function G of the synaptic drive,

$$a_i(\mathbf{x}) = G[e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\Sigma^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i)}],$$

where $\bar{\mathbf{x}}_i$ is the centre of the i^{th} neuron's tuning curve, and Σ is the covariance matrix. G (identical for each neuron) is taken from the model of Gruber et al. [141]. This is a hybrid model in which slow ion channels are modelled in detail, and the firing rate is modelled abstractly as a function of the resulting membrane potential. Figure 7.9B shows the tuning of an example neuron. A single neuron has both "low" and "high" tuning curves, depending on the states of the hysteresis-related channels. The low curve is the narrower one that the neuron exhibits when hysteresis-related channels oppose firing. The high curve is the broader one that the neuron exhibits when these channels promote firing.

A population of these neurons can code progress though a sequence of events, where each event corresponds to a location in the space in which the neurons are tuned. This happens when 1) each neuron's low curve is aligned with an event region, and 2) each neuron's high curve spans all the event regions later in the sequence. This situation is illustrated schematically in Figure 7.10, for a three-event sequence. As the sequence occurs, the neurons that code each event switch to the up state and begin to fire. The sequence is aborted if the state variable leaves the neurons' high curves. Figure 7.10 shows simulations of a network that codes for the sequence $A \rightarrow B \rightarrow C$. A post-synaptic neuron is inhibited when these events occur in order. This example is quite specific, but it illustrates that infinite-dimensional memory does not preclude the possibility of decoding complex functions of input history.

7.5 Firing Dynamics can be Ignored

The averaging code (Chapter 4) essentially ignores variations in the density of tuning curves across the represented space. One consequence is that individual synaptic weights can be learned using only local information, if the post-synaptic neuron is clamped to have the correct activity pattern during learning.

A further consequence is that the code is insensitive to the precise firing rates of the presynaptic neurons. The firing rates determine how much different synapses

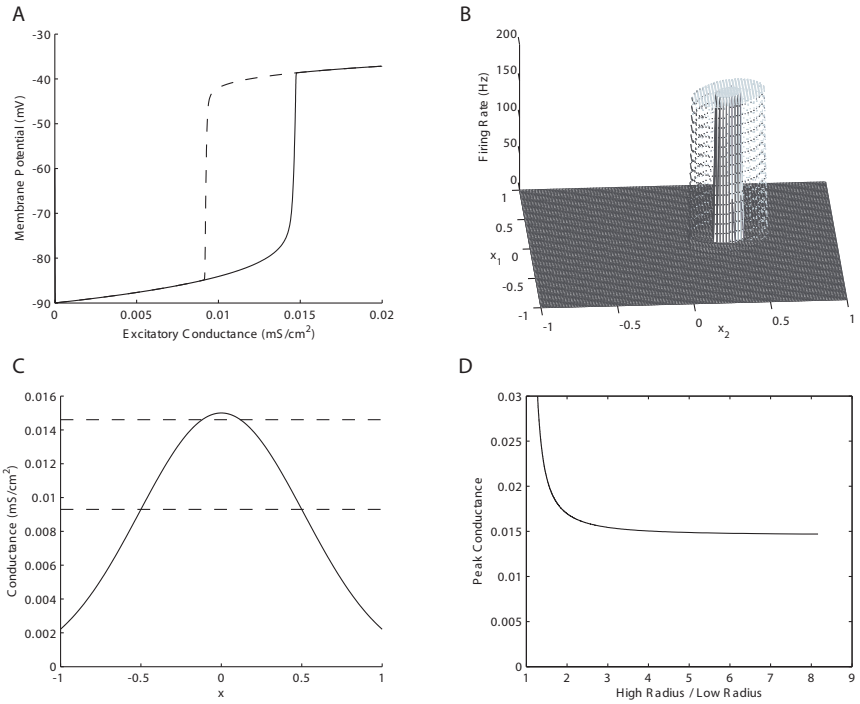

Figure 7.9: Hysteretic tuning curves of medium spiny neurons. A, Hysteresis in membrane potential as a function of excitatory conductance in the model of Gruber et al. [141], with high dopamine concentration. The solid line shows the mean membrane potential as conductance rises from zero, and the dashed line shows the potential as conductance falls from a high value. The neuron's firing rate is a function of this slowly-varying component of the membrane potential. B, An example neuron's low and high tuning curves in a two-dimensional space. The high curve encompasses the narrower low curve. The neuron will only begin firing if x enters the smaller central region, but it will then keep firing as long as x remains within the larger region of the high curve. C, With driving current as a Gaussian function of the represented variable, the height of the Gaussian function determines the relative areas of the high and low curves. This provides a convenient way to construct neuron models with predetermined tuning. The horizontal dashed lines correspond to the onset and offset thresholds, which can be seen in (A). The higher the peak of the Gaussian function, the larger the low curve will be relative to the high curve. D, The peak conductance that corresponds to a range of ratios of low and high radii. Given this ratio, the width of the Gaussian function then determines the absolute radii.

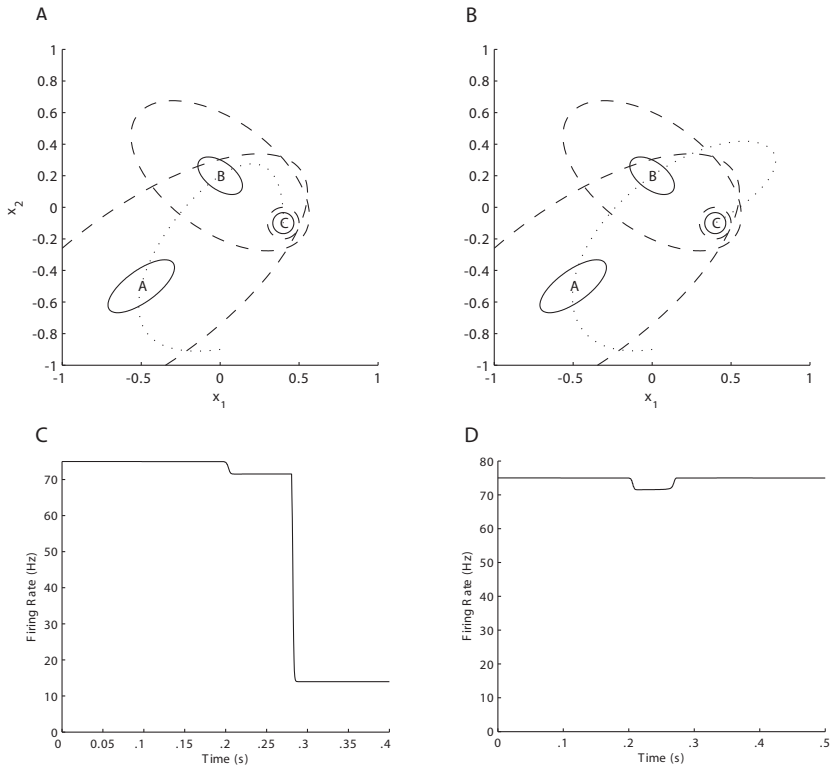

Figure 7.10: Sequence representation by hysteretic neurons. A, Schematic representation of the tuning of three neurons. The solid ellipses show the extents of the neurons' low curves, and the dashed ellipses that encompass them show the extents of the corresponding high curves. The low curves of the neurons are aligned with three regions of interest (labelled A, B, and C). The dotted line shows a trajectory through the state space that passes through these three regions in sequence, without ever leaving the high curves of neurons earlier in the sequence. B, As (A), except that the trajectory leaves the high curves of the A and B neurons before completing the sequence. C, Firing rate of a neuron that receives inhibitory projections from the neurons in (A), as x progresses through its trajectory. The firing rate drops sharply when the sequence is completed. D, The firing rate does not drop when the sequence is interrupted, i.e. when x leaves the high curves in (B). The output neuron in this model is tuned only to the sequence $A \rightarrow B \rightarrow C$; it does not respond for example to the same events in a different order.

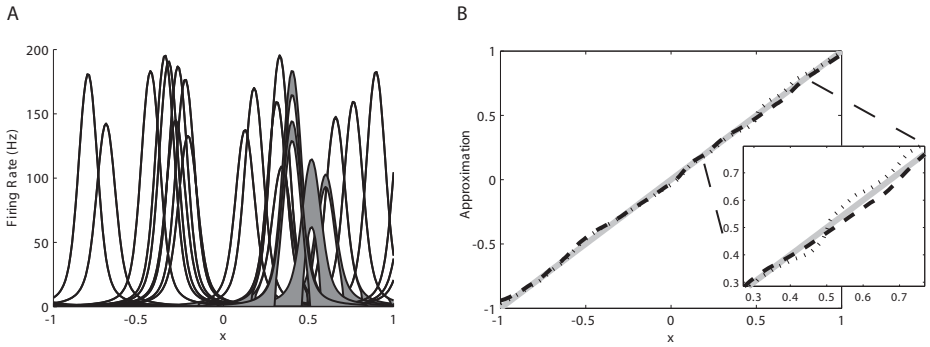

Figure 7.11: Effect of spike-frequency adaptation on the averaging code. A, Tuning curves drawn from a population of 100 adapting LIF neurons. Grey regions indicate the shrinking of some tuning curves due to adaptation, after the population has represented $x_0 = 0.5$ for one second. B, Effect of adaptation on decoding accuracy with independent decoders $\phi_i = f(\mathbf{x}_i^c)$. The gray line indicates the ideal decoding of x, the dashed line indicates the estimate $\hat{x}_{on}$, with onset firing rates. Finally the dotted line indicates the estimate $\hat{x}_{0.5}$, after the population has fully adapted to the represented value 0.5. The error of this estimate is increased around $x = 0.5$.

contribute to a decoded estimate, but if they contribute a little more or less, the estimate is not greatly affected.

Figure 7.11 illustrates this phenomenon with adapting neurons. This figure shows representative tuning curves of an ensemble both before adaptation, and after a value x_0 has been represented for one second, so that neurons tuned near x_0 have adapted to equilibrium. Neurons that are centred closer to x_0 adapt more strongly. The adapted code (with independent decoders $\phi_i = f(\mathbf{x}_i^c)$, i.e. equal to the value of the estimated function at the centre of the neuron's tuning curve) is less accurate around x_0 after adaptation. In particular, it is more sensitive to asymmetry in the tuning curves to the left and right of x_0. Roughly, the *error* has increased in proportion with the decrease in the firing rate of the neurons tuned to x_0. This is very different from the linear code, in which the *estimate* changes in proportion with changes in firing rate.

Other dynamics have similar effects on this code, modulating only the error of the estimate rather than the mean. Figure 7.12 shows an example simulation with intrinsically-bursting neurons. In this network, individual neurons that code the represented value can be completely silent between bursts. However, the neurons burst asynchronously, so that some of the neurons that code each value are active at any given time. Thus the representation carries on fairly smoothly despite the strong underlying oscillations.

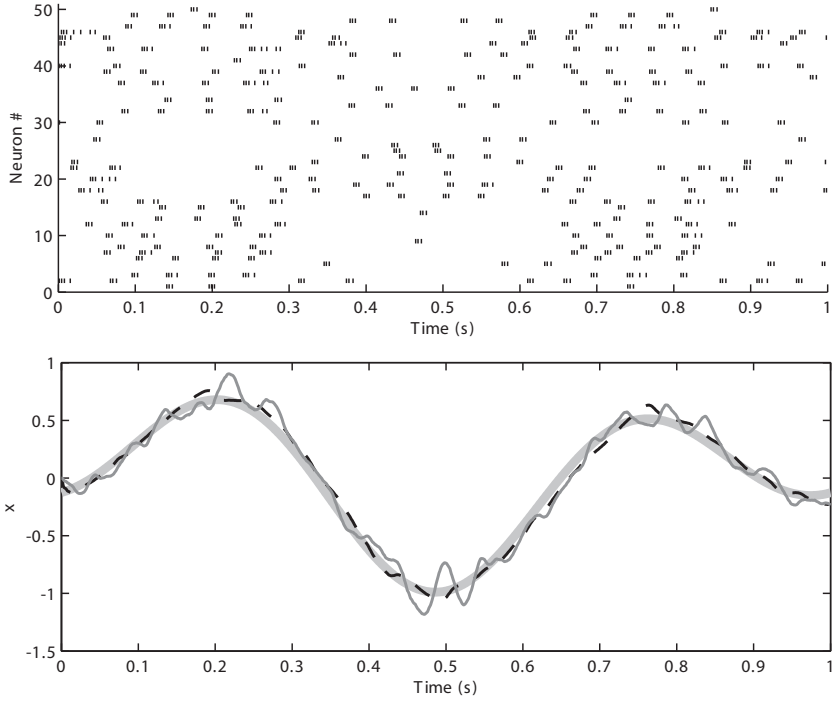

Figure 7.12: Effect of intrinsic bursting on the averaging code. The neurons in this model are bursting pyramidal neurons [198], with driving current that is a Gaussian function of a represented variable x. Each row of hash marks in the top panel indicates the spikes times of a different representative neuron, from a population of 500. Each neuron fires in repeated short bursts when x is near a certain value. The bottom panel shows the ideal value of x (thick solid line) and the linear decoding (thin solid line) and average-based decoding (dashed line) of the bursting activity.

7.6 Discussion

The majority of neurons have intrinsic dynamic processes that affect the time course of the firing rate. This chapter has explored the ways in which the intrinsic dynamics of many cells can combine to influence the dynamics of the population code, and has shown that this relationship is complex.

The NEF takes the important step of integrating population coding and network dynamics, but in doing so, it makes the simplifying assumption that the post-synaptic currents are the dominant source of feedforward ensemble dynamics. Because firing-rate adaptation is common, and because it shares some low-pass properties with PSC dynamics, the first part of this chapter explored whether adaptation might play an analogous role, as dynamical system memory. This is possible, at least in specific circumstances, i.e. when 1) all neurons in the ensemble adapt with the same time constant; 2) the onset responses are linear; and 3) either a) some of the neurons adapt to lesser extents (or not at all), and the high-level network dynamics are not much faster than adaptation, or b) there is a specific relationship between the time constants of PSC and adaptation dynamics. These results begin to incorporate cell-intrinsic dynamics into the coherent view of population coding that the NEF provides. But they do not account for cell-intrinsic nonlinear, oscillatory, or heterogeneous dynamics.

Two different perspectives on the effects of heterogeneous cell dynamics were discussed. First, Section 7.3 showed that heterogeneous linear dynamics can span a set of ensemble transfer functions, with the actual transfer function depending on the synaptic weights with which the ensemble drives post-synaptic neurons. If such an ensemble projects recurrently onto its own neurons, then the resulting network can have a wide range of dynamics, depending on the feedback strength. However, the family of possible dynamics in this case does not necessarily have the same generality as those that can arise from single-time-constant PSC dynamics. From another perspective (Section 7.4), diverse cell dynamics can be interpreted as diverse tuning curves in the space of input history. These tuning curves allow post-synaptic neurons to decode functions of the input history.

Finally, it was shown that cell dynamics have little relevance in the context of the averaging code that was introduced in Chapter 4. This raises the possibility that in some circumstances, instead of contributing strongly to computation, cell dynamics might have comparatively subtle effects on the accuracy of the population code.

7.6.1 Future Work

Computations on input history are probably also enabled by diverse synaptic dynamics, such as short-term synaptic depression and facilitation. It was pointed out in Section 7.2 that synaptic depression (which is present in the striatum [119, 9] and globus pallidus [149]) has an inherently nonlinear effect on network dynamics.

However, this is not at odds with potential roles in the approximation of transfer functions for limited sets of inputs, or in history encoding. There is substantial diversity in synaptic dynamics, both across synapses onto a single neuron, and across synapses from a single neuron – a feature that is conserved from the invertebrate neuromuscular junction [30] to the mammalian cortex [243]. As suggested previously [243, 4], this diversity should provide a rich substrate for dynamic computations.

Dynamics that arise from recurrent micro-circuits might also play roles that are similar to the cell dynamics studied in this chapter. An ensemble of microcircuits could be considered in the same light as an ensemble of neurons, adding considerably to types of unit dynamics that could potentially underlie a dynamic population code.

Finally, while this chapter has focused on the dynamics of spike *rates*, it would also make sense to explore dynamics from the perspective of individual spikes. The commonly-reported spike-triggered average stimulus is the mean pattern of input that causes a spike. It is possible to model a neuron's response as a spike probability, which is a function of the covariance between the input history and the spike-triggered average [113]. This approach could be generalized to multiple dimensions, perhaps using principal components of the inputs leading up to spikes. This should produce dynamic tuning curves in which the spiking probabilities take on more extreme values (i.e. closer to zero or one) as more dimensions are considered (see [57] for a similar approach). Spike-centric dynamic tuning curves should be examined further as substrates of population codes. While the rate-centric tuning curves of this chapter provide high-level descriptions of spiking activity, they treat the high-frequency components of the resulting synaptic currents as noise. Spike-centric tuning curves would not confound spiking activity and noise in this manner.

7.6.2 Relationships with Liquid Computing

There is both a parallel and a potential interaction between history decoding from diverse cell dynamics (Section 7.4) and history decoding from a liquid-state machine [232] (or echo-state network [183]).

A liquid-state machine consists of two parts: 1) a "reservoir", which consists of a recurrent network, and 2) linear read-out neurons. As mentioned in Chapter 5, high-dimensional recurrence in the reservoir leads to diverse dynamic responses to input among the reservoir neurons. The read-out neurons can then decode functions of the input history from the reservoir. This is analogous to the decoding of history from diverse cell dynamics (Section 7.4).

However, reservoir neurons in a liquid-state machine might have intrinsic dynamic properties themselves, which would interact with recurrence to shape the overall dynamics. In this case, the key property of the reservoir, i.e. that it has diverse dynamic responses, would be maintained. Therefore cell dynamics might

have relatively subtle effects on the machine's computational power. This is another case (in addition to the averaging code) in which cell dynamics might have much less impact on computation than their strong influence on cell firing would suggest.

7.6.3 Conclusion

The NEF describes a simple and powerful relationship between cell and network dynamics. In trying to extend this account beyond the dynamics of post-synaptic currents, the original elegance has unfortunately evolved into a patchwork of alternatives and special cases. Given the variety and complexity of firing-rate dynamics, this is not surprising. However, it is interesting that one of these cases (adaptation-based memory) is a generalization of NEF dynamics, and that two other cases (the spanning of transfer functions, and representation of input history) are closely related to the NEF account of static transformation. For this reason, these cases represent an additional link between dynamics and population coding.

Chapter 8

Plasticity and Population Coding

Synaptic plasticity and population coding are two of the key topics in the theory of large neuronal circuits, but the relationships between these topics are not well characterized. This chapter explores the effects of Hebbian synaptic plasticity on population codes. It is shown that Hebbian plasticity can change both the neurons' preferred direction vectors, and the dimension of the coded space. Lateral connections within a population modulate these effects. Specifically, in diverse networks (including principal component analyzers, winner-take-all networks, and self-organizing maps), lateral inhibition consistently increases the dimension of the code, and lateral excitation generally decreases the dimension. Finally, it is shown that cell-instrinsic properties can lead to tuning diversity among neurons with the same preferred direction. These results illustrate how Hebbian plasticity can shape some of the key properties of population codes, such as the dimension and redundancy of tuning curves.

8.1 Introduction

As discussed in previous chapters, computation in neural networks arises from a combination of tuning curves and synaptic weights. In population-coding models, tuning curves are abstract functions of the variables with which the corresponding neural activity varies most clearly (e.g. arm-movement direction). But in reality, this tuning is actually a complex function of upstream synaptic weights. For example, a certain neuron in the visual cortex might be compactly described as an edge detector, but its responses are ultimately a function of the activity of retinal photoreceptors. The main thing that distinguishes this neuron from another neuron with different orientation tuning is the synaptic weights along the path to these neurons, from the photoreceptors. Consequently, although it is convenient to analyse population codes in terms of synaptic weights and tuning curves, the tuning curves are themselves determined by other synaptic weights.

A rough scaffold of synaptic connections is encoded genetically [349]. This genetic scaffold determines the major fibre tracts along which neurons project, as well

as the types of neurons with which a neuron will synapse, and other high-level statistics of these connections. However, the genome, which in humans consists of about 3 billion base pairs, is not nearly rich enough to define the locations and strengths of all the roughly 100 trillion synaptic connections in the human brain. Although the fine structure of the synaptic connections has a major impact on computation, most of the details are established long after conception, through experience-dependent synaptic plasticity. The fine structure that arises from synaptic plasticity is the main factor that distinguishes the brain of an infant from that of an adult.

This chapter considers the population-coding variables of the NEF, such as preferred-direction and decoding vectors, in terms of the effects of plasticity on the synaptic weights. As discussed in previous chapters, the NEF provides a way to model synaptic weights without addressing the question of how they might arise from synaptic plasticity. When a model is developed, the modeler specifies measurable parameters, such as distributions of firing rates, membrane time constants, and preferred directions, as well as the mappings performed within each projection, which can be hypothesized on the basis of electrophysiological or behavioural data. The fine structure of the synaptic weights, which cannot be measured, is then derived as a function of everything else. This approach does not require learning, which is advantageous in that 1) learning can be computationally intensive, and 2) the learning rule that leads to the necessary synaptic weights is not always clear. Thus the NEF approach allows the theorist to focus on instantaneous information-processing properties of a network, without being forced to consider simultaneously the question of how the network might have formed.

On the other hand, the fact that the strength of synapses in the brain is determined by plasticity may constrain the network structure, so models that skip this step are open to criticism. The mechanisms of plasticity are complex and not well understood, either in terms of the underlying molecular mechanisms or in terms of quantitative theory (reviewed by [76, 3]). Given the current state of knowledge, it would be surprising if existing models that learn to perform computations through plasticity did not contain considerable inaccuracies and simplifications. So, a model constrained by learning rules is not necessarily more realistic than one that is instead well-grounded in available experimental data. However, inability to identify a learning rule for a model that is at least plausible (i.e. in which the information used by the rule is probably available at the site of plasticity) is a considerable limitation. Rather than identifying plausible learning rules for specific population-coding models, the goal of this chapter is to describe the role of plasticity in shaping population codes in general.

Current network models of plasticity are highly idealized, and focus on quantifying the consequences of the most basic principles of physiological plasticity. Perhaps the most basic principal is that plasticity at a given synapse is typically a function of activity in both the presynaptic and post-synaptic neurons. This type of plasticity is called Hebbian, due to the early proposal by Hebb [152] that an excitatory synapse should strengthen whenever the presynaptic neuron causes the post-synaptic neuron to fire. Hebbian network models perform unsupervised learn-

ing, which results in some form of implicit representation of the statistics of the input. Despite a great deal of theoretical work on Hebbian plasticity, the role that Hebbian plasticity must play in establishing key characteristics of population codes, such as the distribution of preferred direction vectors, has received little attention. In particular, network models of synaptic plasticity usually assume *small* numbers of *independent*, *high-fidelity* neurons. In contrast, physiological networks contain *large* numbers of *correlated*, and apparently *noisy* neurons (this distinction has been emphasized by C.H. Anderson; personal communication). This chapter considers Hebbian plasticity in the latter context, as the basis of correlated population codes.

8.2 Decomposing Synaptic Weights

Recall from Chapter 3 that the NEF [111] allows one to analytically determine the synaptic weights that calculate a map $f(\mathbf{x})$, in a low-dimensional population code. Specifically, the weights can be found from the preferred direction vectors $\tilde{\phi}_j$ of the post-synaptic neurons, and the vectors $\phi_i^{f(\mathbf{x})}$ that optimally decode $f(\mathbf{x})$ from the presynaptic neurons, as

$$w_{ji} = \alpha_j \tilde{\phi}_j^T \phi_i^{f(\mathbf{x})},$$

where α_j is a scale factor that is common to all the synapses onto the j^{th} post-synaptic neuron.

If the post-synaptic population has dimension d, and the numbers of presynaptic and post-synaptic neurons are m and n, respectively, then the weight matrix W has $d(m+n)$ degrees of freedom. Usually d is much less than both m and n, so that there are many fewer degrees of freedom than synaptic weights. For example, if $d = 3$ and $m = n = 1000$, there are 6x10^3 degrees of freedom among 1x10^6 synaptic weights.

In contrast, plasticity can operate independently on each synaptic weight. Barber [38] pointed out that to relate the potentially high-dimensional weights that arise from plasticity to a low-dimensional population code, singular value decomposition can be performed on the synaptic weight matrix. Singular value decomposition factors a matrix A into a product of three new matrices, $A = USV^T$. S is diagonal, and contains the eigenvalues of $\sqrt{A^T A}$, which are called the singular values of A. These are ordered from largest to smallest. The columns of U and V are orthonormal vectors. The matrix of rank r that is most similar to A (so that the mean-squared difference between matrix elements is minimal) is US_rV^T, where S_r is obtained by setting all but the first r diagonal elements of S to zero. If A does not have full rank, then some of these diagonal elements are zero to begin with.

The rows of U can be interpreted as preferred directions of the neurons in the post-synaptic population. Similarly, the rows of V can be interpreted as decoding vectors of the neurons in the presynaptic population, where each dimension decodes a one-dimensional function of presynaptic activity, over the space that is represented by the presynaptic population. Importantly, an arbitrarily accurate approximation

of synaptic weights can be obtained with lower-dimensional encoders and decoders, by ignoring some of the singular values that are close to zero. For example, if there are only two singular values much larger than zero, the product $U_2 S_2 V_2^T$, where U_2 and V_2 contain only the first two columns of U and V, will provide a very good approximation of W.

Note that regardless of the size n of the presynaptic population, the dimension d of the post-synaptic population is equal to the number of non-zero singular values of W. This is because 1) if $rank(W) < n$, then W projects pre-synaptic signals into a lower-dimensional space, and 2) $rank(W)$ cannot exceed n. If some of the singular values are very small (but not zero), then the corresponding dimensions have correspondingly little influence on the activity of the post-synaptic neurons, and they are probably not relevant to the function of the network.

Figure 8.1 illustrates these correspondences by decomposing a weight matrix that is derived from a product of NEF encoding vectors (preferred direction vectors) and decoding vectors. Note that as described in Chapter 3, it is sometimes convenient to break down a neural computation into two parts: 1) a decoded function $\mathbf{f}(\mathbf{x})$, and 2) a linear transform A. In this case $w_{ji} = \alpha_j \tilde{\phi}_j A \phi_i^{\mathbf{f}}$. The decomposed decoding vectors in this case will correspond to the product $A\phi_i^{\mathbf{f}}$. So clearly, a given V matrix will correspond to any combination of A and $\mathbf{f}(\mathbf{x})$ that give the same product. The fact that this product cannot be uniquely decomposed is not a limitation – in the NEF, the decomposition of a map into A and $\mathbf{f}(\mathbf{x})$ is for analytical convenience, and has no physical meaning.

Barber [38] applied this decomposition to sigmoidal backpropagation networks, and argued that the technique was useful for understanding the computations performed by such networks. Backpropagation networks are valuable engineering tools, in that they can learn complex input/output mappings by generalizing from examples. However, even if they perform correctly, it is often difficult to understand how they make decisions, and they are often criticized as "black-box" models for this reason. Understanding is further frustrated because two backpropagation networks that perform identically may have quite different patterns of synaptic weights. This is because error-driven learning ceases when performance is good, regardless of whether a more elegant or principled solution is possible. This variation can be reduced by constraining the network to have the minimum degrees of freedom needed to perform the task, however this number may not be known in advance, and in any case convergence to a solution that performs well is less likely in a tightly constrained network. Barber essentially argued that it is more practical to extract the low-dimensional structure of the solution afterwards, independently of the learning process.

This study established a new connection between population-coding models and artificial neural networks. However, its physiological relevance was limited by 1) the physiological implausibility of backpropagation, and 2) the questionable relevance of the independent, high-fidelity neurons that make up a feedforward sigmoidal model for understanding the brain. In contrast, the remainder of this chapter

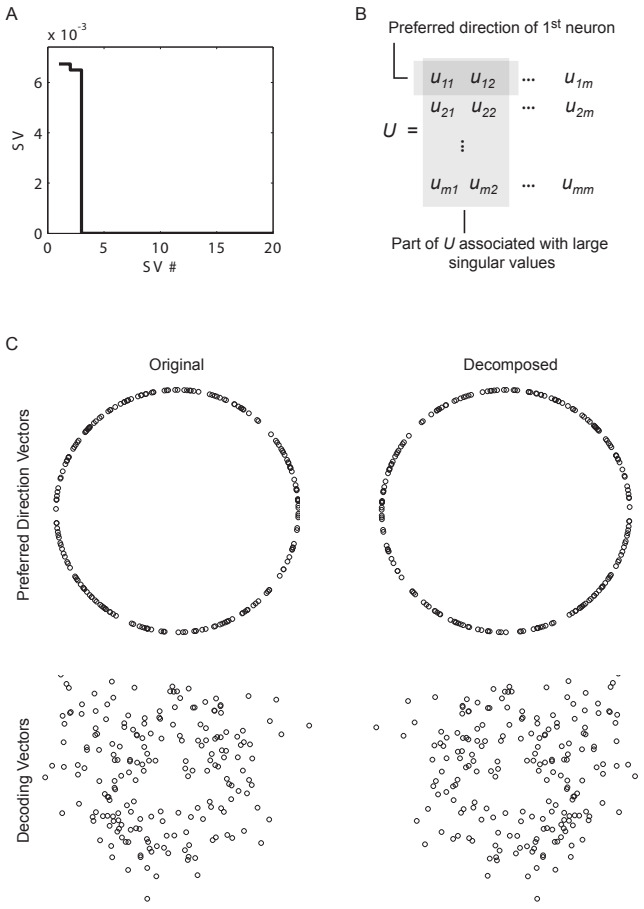

Figure 8.1: Decomposition of NEF synaptic weights into preferred direction vectors and decoding vectors. Optimal synaptic weights for a 2D communication channel between LIF populations were found using NEF methods (Chapter 3), as the product of the preferred direction vectors of the post-synaptic neurons, and optimal decoding vectors of the presynaptic neurons. Singular value decomposition was then performed on the resulting synaptic weight matrix W, to give $W = USV^T$. A, The resulting weight matrix has two large singular values, reflecting the fact that the post-synaptic code is two-dimensional. B, The two-dimensional preferred direction vectors of neurons in the post-synaptic population correspond to rows of a sub-matrix of U that consists of the first two columns. Analogously, the decoding vectors correspond to columns of a submatrix of V^T that consists of the first two rows. C, The preferred-direction and decoding vectors found by decomposing the synaptic weight matrix are the same as the original vectors, except that in general they may be rotated, flipped, and/or rescaled. These changes do not affect the meaning of the code (just arbitrary features of the axis labels). In this example, the original and decomposed vectors are mirror images of each other.

uses the same decomposition method to explore how Hebbian plasticity (which is ubiquitous in neural systems) can establish and shape a biologically-realistic, redundant population code.

However, before turning to Hebbian plasticity in particular, some preliminary observations can be made about the effects that synaptic plasticity can have on a population code, and on the functions calculated using the code. As illustrated in Figure 8.2 (with a supervised-learning network), synaptic plasticity can potentially change: 1) the function that is calculated by a projection; 2) the preferred directions of the post-synaptic neurons; and perhaps most importantly, 3) the dimension of the transform. The dimension of the transform bounds the dimension of the post-synaptic neurons' tuning curves. For example, if the presynaptic population is ten-dimensional, synaptic weights that are two-dimensional will project the presynaptic activity into a two-dimensional space. On the other hand, if the presynaptic population is two-dimensional, ten-dimensional weights will result in a ten-dimensional post-synaptic code, and different represented values will belong to a two-dimensional manifold.

The following sections show that Hebbian plasticity can influence the dimension of a synaptic weight matrix, and also the distibution of the preferred directions of the post-synaptic neurons. It is later argued that either supervised learning or reinforcement learning is needed for flexible modulation of the decoding vectors.

8.3 Hebbian Plasticity

Hebb [152] proposed that if a neuron A causes a neuron B to fire, the synapse from A onto B will strengthen. In the broadest sense, the term "Hebbian plasticity" is now taken to mean any form of plasticity in which changes in synaptic strength depend on presynaptic activity, post-synaptic activity, and other information that is available locally at the synapse (e.g. the synaptic weight itself).

There is now a great deal of experimental evidence for Hebbian plasticity in this broad sense [3], and for more specific variations. One variation that has received a great deal of attention recently is spike-timing-dependent plasticity (STDP), in which the synapse is strengthened when the presynaptic neuron fires a few milliseconds before the post-synaptic neuron, but weakened when post-synaptic neuron fires first. This variation is closer to Hebb's original proposal than temporally-symmetric variations. STDP has been observed in the cortex [242] and recently in the striatum [296]. The theoretical BCM rule (after Bienenstock, Cooper & Munro) encapsulates another variation for which there is recent experimental support (see [78]). In this rule, the change in synaptic weight is a product of the presynaptic firing rate and a nonlinear function of the post-synaptic firing rate. This nonlinearity causes the synaptic weight to decrease when post-synaptic activity is lower than some threshold, and to increase when it is higher. The threshold evolves as a super-linear function of the mean post-synaptic firing rate. This stabilizes the neuron's activity so that it does not grow without bound.

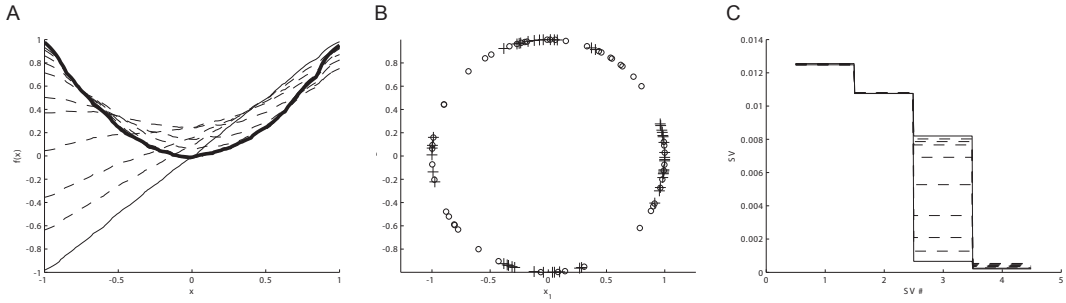

Figure 8.2: Synaptic plasticity can affect the calculation performed by a neuronal projection, the post-synaptic neurons' preferred directions, and the dimension of the post-synaptic population. These effects are illustrated here in a model of a projection from one LIF population to another. The synaptic weights in this model are modified by the supervised learning rule $\Delta w_{ji} = \kappa a_i e_j$, where κ is a learning rate, a_i is the firing rate of the i^{th} presynaptic neuron, and e_j is the difference between the actual firing rate of the j^{th} post-synaptic neuron, and the ideal firing rate (i.e. what the firing rate would be if the projection calculated the target function, *etc.*) The model is trained by presenting it with randomly-selected values of the presynaptic variable x. A, Modulation of decoded function from $f(x) = x$ to $f(x) = x^2$. The thick solid line indicates the least-squares optimal decoding of x^2 from presynaptic activity. The thin solid lines indicate the actual decoding at the start and end of training, and the dashed lines indicate intermediate functions during training. B, Modulation of preferred directions, from uniformly-distributed to clustered around the axes. The circles indicate the preferred directions at the start of training, and the plus-marks indicate preferred directions at the end of training. C, Modulation of the dimension of the post-synaptic code, from two to three-dimensional. The magnitudes of the first three singular values are shown. Initially (thin solid line) there are two large singular values, corresponding to a two-dimensional post-synaptic code. The dashed lines indicate singular value magnitudes at different times during training. Eventually they converge on those of the target mapping (thick solid line). Note that as each feature of the projection is modified by synaptic plasticity, the other features remain constant, e.g. the encoders and the dimension remain constant in (A).

Hebbian networks self-organize to represent their input in a different manner, which is determined by the statistics of the input. In the most basic form of Hebbian plasticity, the synaptic weight w_i from the i^{th} presynaptic neuron onto a post-synaptic neuron varies as

$$\Delta w_i = \kappa a_i b,$$

where Δw_i is the change in the synaptic weight after presentation of a single input,[1] a_i is the firing rate of the i^{th} presynaptic neuron, b is the firing rate of the post-synaptic neuron, and κ is a learning-rate constant that scales the rate of change of the weights. Slower plasticity can prevent a network from over-specializing on the basis of a small sample of inputs.

Because stronger correlation between pre- and post-synaptic activity leads to stronger weights, and *vice versa*, this simple scheme causes synaptic weights to grow without bound. However, different weights do not grow equally. Instead, the weight vector grows most quickly along the first principal component of the variation in the presynaptic activity **a** [92].

Oja [281] proposed a variation on this rule in which different synapses compete for resources, so that the norm of the synaptic weight vector remains constant over time. More recent experimental evidence indicates that receptors diffuse rapidly in and out of post-synaptic membrane specializations, so that receptor concentration at a synapse (a key determinant of its strength) is in dynamic equilibium [367]. If synapses dynamically attract receptors, rather than fixing them in a molecular scaffold (the classic view), this would be consistent with the competition between synapses that Oja suggested for theoretical reasons.

In Oja's model, if the post-synaptic neuron has a linear response function, and the learning rate is low, then synaptic plasticity can be expressed as

$$\Delta w_i = \kappa b(a_i - w_i b). \tag{8.1}$$

The weight vector does not lengthen over time, but it gradually aligns with the first principal component of the presynaptic activity. The activity of the post-synaptic neuron then reflects the projection of presynaptic activity onto this principal component. This simple Hebbian network therefore models key statistical information about its inputs.

A key point of divergence between Oja's model and biological neurons is Oja's assumption that neuron response functions are linear, in contrast with the rectification of physiological firing rates at zero, and their saturation at high values. However, Oja's derivation is easily adapted to the nonlinear case (see Appendix), yielding

$$\Delta w_i = \kappa b(a_i - w_i \sum_i w_i a_i). \tag{8.2}$$

[1]For computational efficiency, models of synaptic plasticity usually ignore changes in neuron activity over short time scales, and assume discrete trials during which the activity of each neuron is constant.

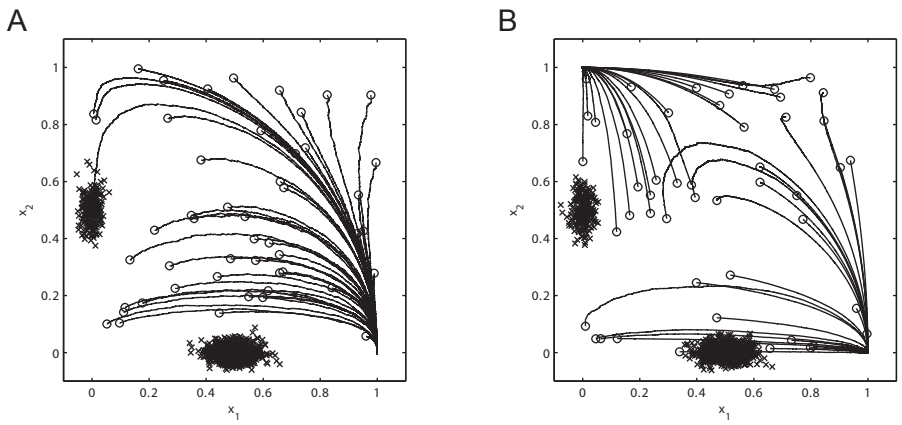

Figure 8.3: Nonlinear Oja neurons do not necessarily extract the first principal component of the input. This is illustrated in a network with two inputs, and multiple independent output neurons. Synaptic weights have random initial values, and are plastic according to the Oja rule. A, Linear output neurons. The inputs are drawn from a bimodal distribution, which is the average of two Gaussian functions. Random input samples from this distribution are marked as x's. The lower-right peak has higher density, so that the first principal component points along the x_1 axis. The circles indicate random initial weight vectors of different post-synaptic neurons, and the lines that begin at each circle show the trajectories of these weight vectors over the course of training. All of the weight vectors align with the first principal component of the input. B, Nonlinear output neurons. The inputs in this simulation are drawn from the same distribution as in (A), but in this case, a neuron's weight vector may align with either the first or second principal component, depending on its initial value.

A nonlinear Oja neuron can also extract the first principal of the input. However, depending on the structure of the input, it may extract something different. Figure 8.3 shows an example in which different nonlinear Oja neurons that are driven with the same input extract the principal components of two different regions of an input with a bimodal probability distribution.

Bounds on the magnitudes of the synaptic weights can have a similar effect [92], so that two different post-synaptic neurons extract different features of the same input.

8.4 Dimension Control by Lateral Connections

As discussed above, if synaptic strength in a feedforward network is established by the Oja rule, then the post-synaptic neurons will all become tuned to the first

principal component of the input (or due to nonlinearities, perhaps to one of a few other statistical features of the input). In this case, the rows of the synaptic weight matrix will be parallel, and the matrix will have rank one. The post-synaptic population will be one-dimensional, and will represent the first principal component of the input.

Lateral connections among neurons in the post-synaptic population can change this behaviour dramatically. This section discusses the effects of lateral connections on the dimension of the feedforward weight matrix, in several well-known network types.

There are many variations on the theme of Hebbian learning with lateral connections, and it is not possible to address all of them here. However, two network types are of particular interest. The first is principal component analyser networks, which extend the Oja model to include lateral inhibition. Section 8.4.1 makes the obvious but important point that the sparseness of lateral connectivity in principal component analysers impacts the dimension of the population code. The second important type of laterally-connected Hebbian network is the self-organizing map (SOM). Self-organizing maps are of interest because they model the topological organization of neural tuning that is observed in many brain areas. Section 8.4.3 shows how spatial patterns of lateral connectivity in these networks (which may be genetically-determined, in the physiological circuits they model) influence the dimension of the code, suggesting a link between population coding and spatial network properties. Additionally, section 8.4.2 discusses Hebbian learning in winner-take-all networks. These networks have similarities with both principal component analysers and SOMs, making it clear that all of these networks types belong to a larger family. Notably, principal component analysers [37], self-organizing maps [300], and winner-take-all networks [351] have all been proposed as models of the striatum.

In general, it will be shown that inhibitory lateral connections in these networks increase the dimension of the code, and excitatory lateral connections reduce the dimension.

8.4.1 Principal Component Analysers

Oja's original model has been extended to extract multiple principal components, by introducing lateral inhibitory connections between the post-synaptic neurons. In one well-known variation on this theme, by Kung & Diamantaras [212], the matrix of inhibitory lateral weights is lower-triangular. The first post-synaptic neuron receives no lateral inhibition, so it behaves like an Oja neuron, and converges on the first principal component of the input. The second post-synaptic neuron is inhibited by the first. This inhibition strengthens when the firing rates of the first and second neurons are correlated. The result is that the second neuron does not align with the first principal component, but rather the second (the largest source of variance that is orthogonal to the first). The third neuron is inhibited by the

first two, and aligns with the third principal component, and so on. In another variation [121], the post-synaptic neurons are connected symmetrically. In this case, correlated neurons are also driven apart by mutual inhibition, but there is no hierarchy. Consequently, m post-synaptic neurons do not align precisely with the first m principal components. But they still span the m-dimensional principal subspace.

Contrasting these networks with a post-synaptic population of independent Oja neurons makes it clear that the effect of lateral inhibition is to increase the dimension of the feedforward weights. As shown in Figure 8.4, the dimension of the feedforward weights in these networks is the minimum of 1) the dimension of the *variance* in the input pattern (which may be less than the number of input neurons), and 2) the number of post-synaptic neurons.

To return to population coding, redundancy can easily be introduced in these networks through alternative pattens of lateral connectivity. For example, in an asymmetric Kung & Diamantaras [212] network, if some of the neurons do not receive lateral inhibition from the first, their weight vectors will align with those of the first neuron. In general, sparser lateral connection matrices will allow greater redundancy. Figure 8.4C illustrates a case in which eight post-synaptic neurons form two completely independent groups, so that the resulting code is at most four-dimensional.

8.4.2 Winner-Take-All Networks

In a winner-take-all (WTA) network, only the post-synaptic neuron that is driven most strongly is active at any given time. This behaviour is often modelled phenomenologically. But in a more detailed model, it can arise as a result of strong mutual inhibition among the post-synaptic neurons. WTA networks (even the phenomenological ones) therefore behave like networks with strong lateral inhibition. Accordingly, Hebbian learning in the feedforward weights of such a network leads to a high-dimensional code, as shown below.

Simple Winner-Take-All Model

Before turning to Hebbian learning in WTA networks, this section introduces a simple WTA model based on lateral inhibition, which has many similarities with the principal component analysers of the previous section. In this model, the firing rate b_j of the j^{th} post-synaptic neuron is

$$b_j = G[\sum_i w_{ji} a_i + \sum_k q_{jk} b_k + J^{bias}], \tag{8.3}$$

where a_i is the firing rate of the i^{th} presynaptic neuron, w_{ji} is the weight of the feedforward synapse from the i^{th} presynaptic neuron onto the j^{th} post-synaptic

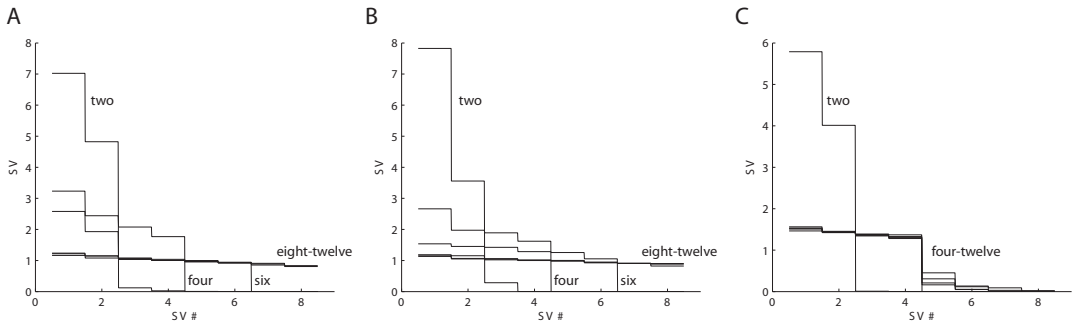

Figure 8.4: Dimension of feedforward weight matrices in principal component analysers with different patterns of lateral connectivity. Forward weights in these networks obey the Oja rule, and lateral weights obey a modified Oja rule, in which mutual inhibition strengthens when the activity of two neurons is correlated (this is sometimes called "anti-Hebbian" plasticity, because the numerical value of the weight becomes lower rather than higher with correlated activity). Each line indicates the magnitude of singular values in a network with 12 input and 8 output neurons. Inputs are drawn randomly from a Gaussian distribution with varying dimension. The labels beside each line indicate the dimension of this variance, within the 12-dimensional space of presynaptic activity. A, Asymmetric lateral connections [212] lead to a code with the same dimension as the variance in the input, up to the number of post-synaptic neurons. The number of large singular values of the feedforward weight matrix corresponds to the dimension of the input variance, up to a maximum of 8 (the number of output neurons). B, Symmetric lateral connections have essentially the same effects on the dimension of the code as asymmetric lateral connections. C, Sparse lateral connections can lead to codes of reduced dimension. In this example, lateral connections are symmetric, but the neurons form two mutually-independent groups of four neurons each. Each group of neurons finds (independently) the four-dimensional principal subspace. Consequently, the resulting code has at most four dimensions, regardless of the pattern of input.

neuron, q_{jk} is the weight of the lateral synapse from the k^{th} onto the j^{th} post-synaptic neuron, J^{bias} is a constant intrinsic bias current, and $G[\bullet]$ is the leaky-integrate-and-fire (LIF) response function (Chapter 3).

The lateral weight matrix Q is static. The off-diagonal entries are negative, modelling lateral inhibition. The diagonal entries are positive, modelling recurrent excitation of each neuron.[2] Figure 8.5 illustrates how a neuron's firing rate decays to zero when the net input is low, and to a high value when the net input increases beyond a threshold.

The forward weight matrix $W = [w_{ji}]$ is established by the Oja rule. In a given trial, differences between the rows of this matrix result in differences in the total forward input to each neuron. The neuron with the weight vector that is most closely aligned with the input vector wins the WTA competition. This model is closely related to the principal component analyzers discussed in the previous section. The main differences are that the WTA model has strong, static lateral weights, self-excitation, and nonlinear response functions.

Additionally, a homeostatic mechanism is added, so that the weights of non-active (non-winning) neurons grow uniformly. A non-winning neuron's weights continue to grow until it wins enough competitions that the normalizing effect of the Oja rule balances continued weight growth. This mechanism forces all neurons to participate in the code (Figure 8.6).

Dimension of Feedforward Weights

The dimension of the WTA network is similar to that of a principal component analyzer with the same structure (Figure 8.7). One interesting difference is that while the dimension of the synaptic weights in the principal component analysers is at most m, the number of output neurons, the dimension of the WTA network is at most $m - 1$. This is because each WTA unit eventually points close to the centre of its winning territory. This forces the preferred directions of different WTA units apart. So for example, with Gaussian input, two units will point in opposite directions in a two-dimensional space, three units will point in directions that are close to co-planar with (0,0), etc. In contrast, lateral interactions in the principal component analyzers tend to make the preferred directions of different neurons roughly orthogonal.

Preferred Directions

In the WTA network with Hebbian feedforward weights, preferred direction vectors point in useful directions.

[2]Differences in forward input can be amplified by lateral inhibition alone, but all-or-nothing competition requires positive feedback.

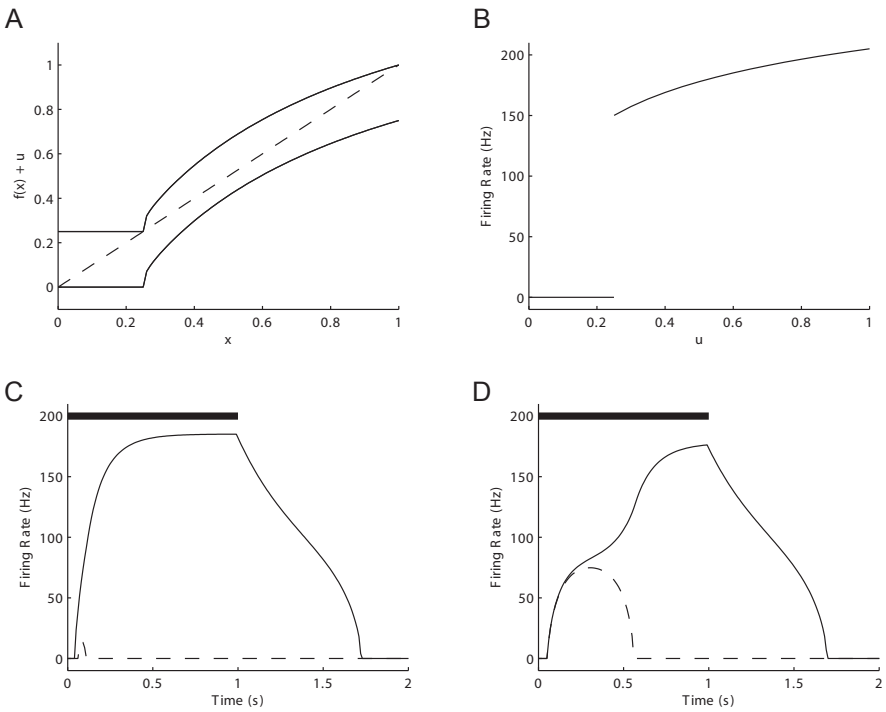

Figure 8.5: Winner-take-all model. A, Stable firing rates with self-excitation. The dashed line $f(x) = x$ indicates the degree of self-excitation necessary to balance exponential post-synaptic current decay, in order to exactly maintain the firing rate at a constant value. The lower solid line shows a scaled LIF tuning curve that corresponds to the degree of self-excitation with zero input. This self-excitation is always below the dashed line, so the firing rate decays to zero from any initial value. The upper solid line indicates self-excitation *plus* input that is just sufficient to activate the neuron. The total input to the neuron is above the dashed line for low firing rates, so that the neuron's firing rate increases over time until the total input crosses below the dashed line at a higher represented value $x = 1$. B, Bifurcation of the firing rate as a function of net input u. The winning neuron must inhibit the others strongly enough that their net input does not cross the bifurcation point (0.25 in this model). C, A competition between two neurons with forward input 0.5 and 0.6 is quickly resolved. The bar indicates the time during which the input is presented. D, A competition between the same neurons with forward inputs 0.5 and 0.501. In this closer competition, it takes longer to clearly establish the winner.

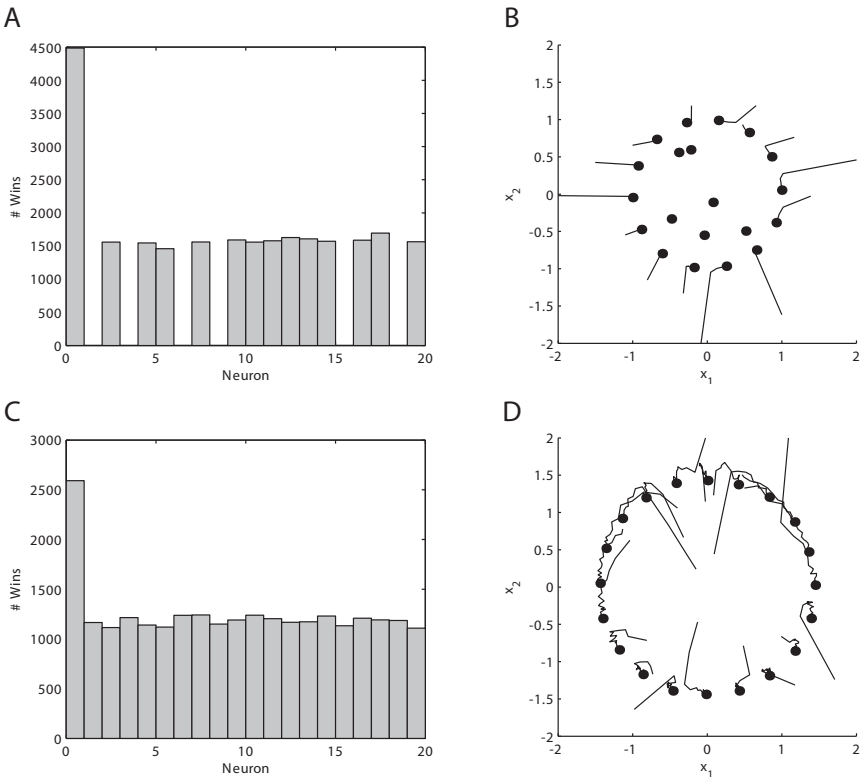

Figure 8.6: Homeostatic weight growth in a WTA network forces all neurons to participate. This is illustrated in a network with two-dimensional Gaussian-distributed input, and 20 output units. A&B, When feedforward weights are modified according to Oja rule alone, some of the neurons never win a trial. A, Histogram of wins by different neurons in the last half a training simulation. B, Trajectories of the weight vectors, which converge on final locations that are marked with dots. The weights of the non-winning neurons are not modified during training. C&D, Here the weight vectors lengthen slightly with each non-winning trial. Neurons that never win are therefore driven more and more strongly until they are competitive, at which point Oja normalization counterbalances weight growth in non-winning trials.

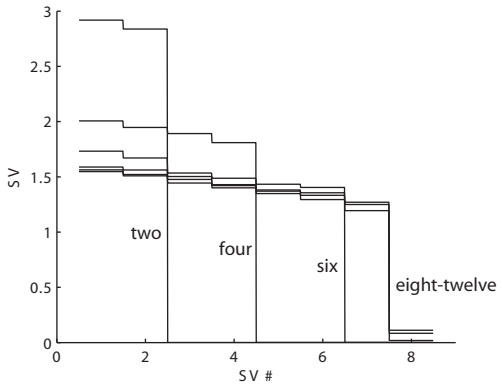

Figure 8.7: Dimension of feedforward weight matrices in a WTA network is similar to that in principal component analysers (see Figure 8.4). The main difference in this case is that the maximum dimension of the code is one less than the number of post-synaptic neurons, as discussed in the text.

The distribution of preferred direction vectors in a population code influences both the accuracy with which different values are represented, and the computations that can be performed on the represented values. In the absence of noise, represented values are unambiguous, as long as the preferred directions span the space. But if the preferred directions are clustered together, then independent noise in the firing of different neurons will introduce relatively less ambiguity in the directions of the clusters, and more ambiguity in other directions. In terms of computation, if neurons are clustered around a small number of preferred directions $\tilde{\phi}_{1..l}$, then functions of the form $f(\mathbf{x}) = f_1(\tilde{\phi}_1) + ... + f_n(\tilde{\phi}_l)$ can be linearly decoded from the neurons' activities, assuming tuning curves along each direction are diverse. Greater diversity of preferred directions leads to more terms in this sum, and simultaneously (assuming a constant total number of neurons) less accuracy in the decoding of each term.

The Hebbian WTA network described above tends to align *weight* vectors with patterns of *neural activity* that are encountered frequently. From a population-coding perspective, this is equivalent to aligning the *preferred direction* vectors with values that are frequently *represented* (Figure 8.8). This means that frequently-encountered inputs will be represented by a greater density of neurons. Thus the code is biased both to minimize the effects of noise and to maximize computational flexibility for values that are represented most frequently. In contrast, the principal component analysers discussed in the previous section do not cluster inputs along important directions in the coded space. For example, in the model of Kung & Diamantaras, different neurons are always aligned orthogonally, regardless of the distribution of the inputs.

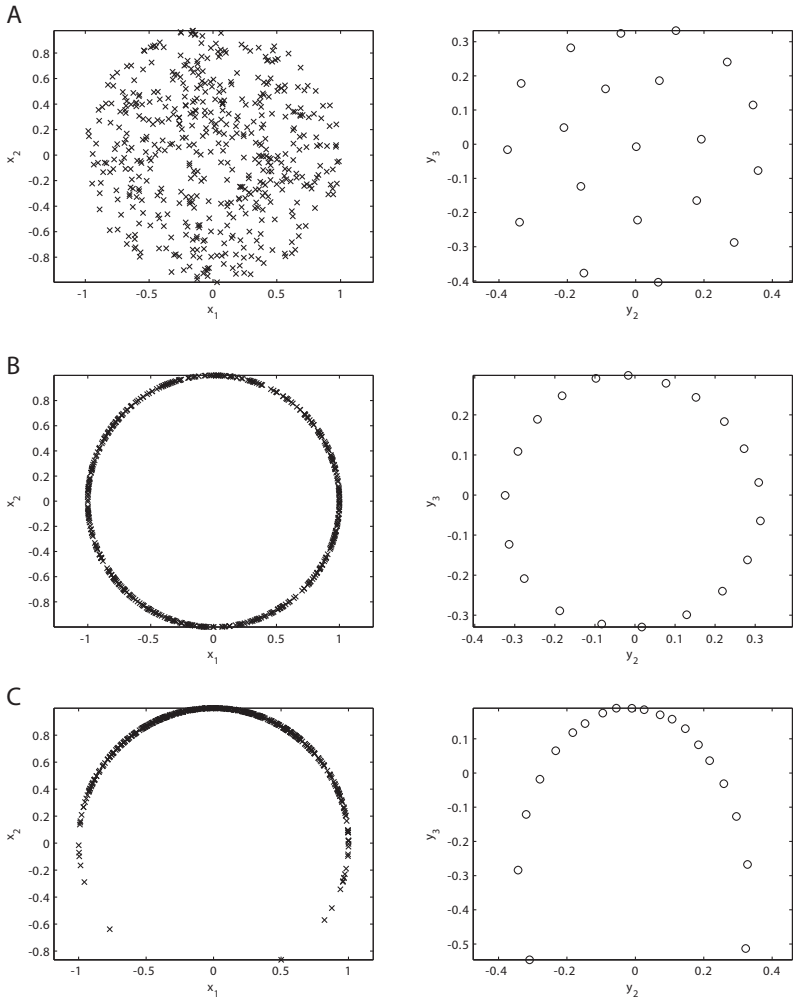

Figure 8.8: Preferred-direction vectors cluster around represented values in a WTA network. The model simulated here consists of fifty two-dimensional presynaptic LIF neurons, and twenty post-synaptic WTA neurons. The left panels show samples from different distributions of values represented by the presynaptic population. The right panels show the preferred directions of the post-synaptic neurons (circles) after training of the network. The preferred directions of these neurons are essentially three-dimensional, and are plotted in their second and third dimensions (they are roughly uniform in the first dimension, reflecting the constant first principal component of the input due to firing rate rectification). Presynaptic represented values are drawn uniformly from the interior of the unit circle (A), drawn uniformly from the edge of the unit circle (B), and drawn non-uniformly from the edge of the unit circle (C; here angles have a Gaussian distribution around $\pi/2$). In each case, the density of the preferred directions reflects the density of the input.

8.4.3 Self-Organizing Maps

A self-organizing map [208] is a type of unsupervised-learning network in which neurons are arranged in a (usually) two-dimensional sheet. Nearby neurons are mutually excitatory, and neurons that are farther apart are mutually inhibitory. After learning, the network forms a topological representation of its inputs, i.e. neurons that are close to each other are similarly tuned. Self-organizing maps (SOMs) have practical applications in the visualization of high-dimensional data. They are also of interest as neural circuit models, because their topological representation of input resembles the topological organization that is common in the neocortex (which is also essentially a thin sheet of neurons) and in other subcortical structures.

A SOM is initialized either with random forward weights, or with forward weights that are distributed to reflect basic statistical properties of the input. A learning trial consists of several stages, in which lateral excitatory and inhibitory interactions are modelled abstractly. First, the post-synaptic neurons participate in a phenomenological winner-take-all competition, so that

$$b_j = \begin{cases} 1 & \text{if } d_j = \sum_i w_{ji} a_i > d_k \forall k \neq j \\ 0 & \text{otherwise} \end{cases} \tag{8.4}$$

where b_j is the activity of the j^{th} post-synaptic neuron, d_j is its total forward synaptic drive, a_i is the firing rate of the i^{th} presynaptic neuron, and w_{ji} are the feed-forward weights. Next, the winning neuron excites nearby neurons, so that their activity levels vary with the physical distance from the winning neuron, according to a kernel function. The kernel is typically unimodal (e.g. Gaussian) with zero mean. These two steps (i.e. WTA and lateral excitation) model lateral inhibition and excitation in a highly idealized but computationally efficient manner. Finally, the forward weights w_{ji} are updated so that the weight vectors of each *active* neuron approach the input pattern, i.e.

$$\Delta \mathbf{w}_j = \kappa b_j (\mathbf{a} - \mathbf{w}_j), \tag{8.5}$$

where κ is a learning rate. This is a form of Hebbian learning.

Kernel Width and Dimension

As in the WTA networks of the previous section, preferred direction vectors in a SOM cluster around frequently-coded values. However, excitatory lateral interactions correlate the activities of nearby neurons. One result of these lateral interactions is that if inputs belong to clusters, some post-synaptic neurons can become tuned to directions between clusters (i.e. if they are close to neurons in each cluster). These lateral interactions can also act like a spatial filter, so that outliers in the input are ignored (Figure 8.9).

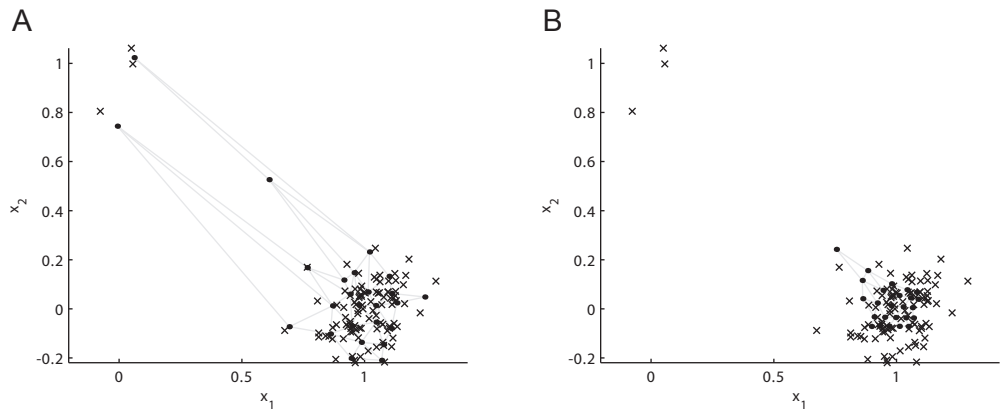

Figure 8.9: Excitatory lateral interactions filter outliers. This is illustrated in a network with two presynaptic neurons and a 5x5 hexagonal sheet of self-organizing output neurons. Sample inputs are shown as x's. Weight vectors of the post-synaptic units are shown as dots, with gray lines connecting units that are adjacent in the sheet (i.e. mutual distance=1). The Gaussian excitatory kernel widths are 0.35 units (A) and 1.5 units (B). With the wider kernel, none of the output neurons align with the outlying inputs near (0,1).

This filtering process can also lower the dimension of the learned feedforward weights (Figure 8.10). This happens when there are dimensions along which the inputs fluctuate weakly, in a manner that is uncorrelated with fluctuations in dimensions of higher variance. In this case, neighbouring neurons are drawn along the direction of high variance, and the low-pass effect of the kernel prevents the preferred directions from diverging along the low-variance direction. However, this smoothing has little effect on the dimension of the forward weights if the low-variance features of the input have a smooth relationship with the high-variance features.

In summary, the winner-take-all character of the self-organizing map can produce a high-dimensional code if there is high-dimensional structure in the inputs. However, the smoothing effect of a broad excitatory kernel can supress features of the input that are 1) weak, and 2) lacking a smooth structure with respect to other features.

8.4.4 Tuning Curves in Laterally-Connected Populations

The previous sections illustrate that lateral connections within a population influence both the direction and the dimension of the neurons' preferred-direction vectors. In the models of previous chapters, a neuron's preferred direction and its intrinsic response function uniquely defined the neuron's tuning curve. However, in

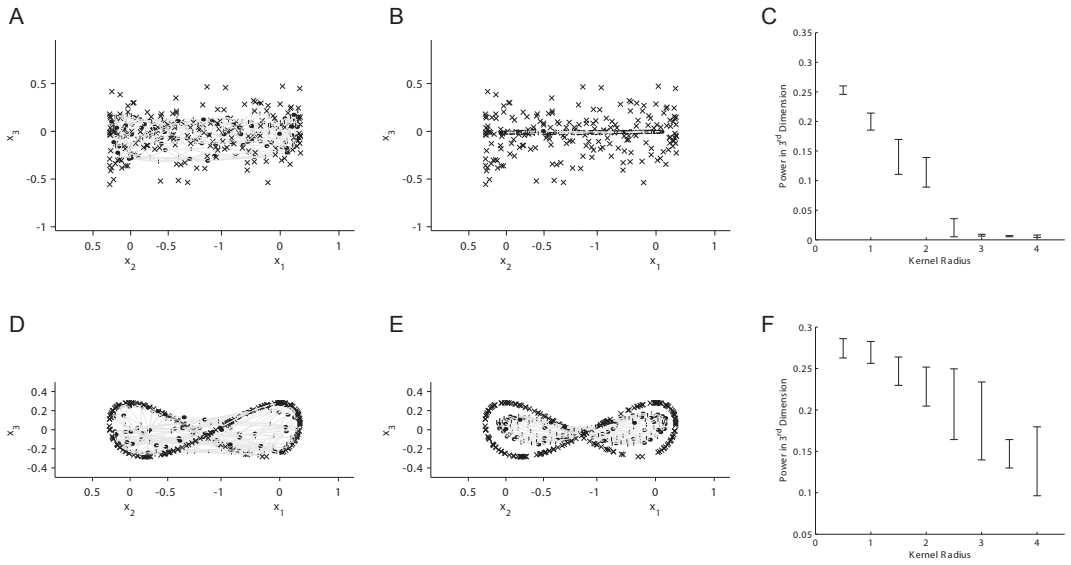

Figure 8.10: A broader excitatory kernel can decrease the dimension of the forward weights. The network shown here consists of three presynaptic neurons, and an 11x14 hexagonal sheet of output neurons. In the top panels, the input consists of a 2D circle with high-frequency noise in the third dimension. A, Looking at the circle from the side, the x's indicate sample inputs, and the dots indicate the three-dimensional weight vectors of the output neurons, after training with a narrow excitatory kernel (1 unit). Weight vectors of adjacent neurons in the sheet are connected by gray lines. B, With a broader kernel (3 units), the high-frequency variations in the third dimension are ignored, and the weight vectors become two-dimensional. C, Mean +/- standard deviation of the magnitude of the third singular value of the weight matrix, divided by the mean magnitude of the first two singular values, as a function of kernel width (5 randomly-initialized networks per kernel width). When the kernel width exceeds 2 neurons, the post-synaptic neurons do not encode the third dimension of the input. D-F, As A-C except that values in the third dimension are a smooth function of values in the other two dimensions. In this case, the post-synaptic neurons represent all three dimensions of the input, even with broader excitatory kernels.

a laterally-connected population, the lateral connections also exert a direct influence on the tuning curves. Of course, the tuning curves still exist within the space that is defined by the dimension of the feedforward weights. That is, regardless of the rank of the lateral weight matrix, the tuning curves have the dimension of the forward weight matrix. However, the shape of each tuning curve is modulated by other neurons in the population.

Figure 8.11 illustrates the effects of mutual excitation and inhibition on the tuning curves of LIF neurons. Lateral interactions incline the tuning curves toward or away from those of other neurons.

These tuning curves also vividly illustrate why lateral excitation and inhibition generally decrease and increase the dimension of the code, respectively. Mutual excitation *forces* neurons to respond strongly to the same input, so that Hebbian plasticity draws the weight vectors together. Mutual inhibition *prevents* neurons from responding strongly to the same input, so that Hebbian plasticity draws the weight vectors apart.

8.5 Diverse Tuning

As shown above, sparse lateral inhibition (or lateral excitation) allows multiple neurons to have similar preferred directions. This leads to redundancy in the neural code, which can potentially reduce the effects of noise. However, redundancy alone does little to aid computation. Recall from Chapter 3 that diversity of tuning curves along the same preferred direction enables linear decoding of diverse functions. How can redundant neurons with shared input and Hebbian plasticity exhibit diverse but correlated responses? One might expect that such differences could arise from intrinsic differences between neurons. This section introduces an extension to the Oja rule that allows for such differences, and leads to diverse tuning.

The extention to the Oja rule consists of including a homeostatic mechanism that regulates the neuron's mean level of activity, through changes in bias current. Physiologically, the intracellular calcium concentration (which increases with firing rate) has been proposed to drive homeostatic regulation of ion channel density (see reviews by [241, 89]). The present extension to the Oja rule is a highly simplified model of homeostasis, which acts on parameters of the leaky-integrate-and-fire (LIF) model, rather than on models of specific conductances.

In the extended model, the firing rate b_j of the j^{th} post-synaptic neuron is

$$b_j = G[\sum_i w_{ji} a_i + J_j^{bias}], \tag{8.6}$$

where a_i is the firing rate of the i^{th} presynaptic neuron, w_{ji} is the corresponding synaptic weight, and $G[\bullet]$ is the LIF response function. J_j^{bias} is an intrinsic bias current. The magnitude of this current varies in such a way that the mean firing

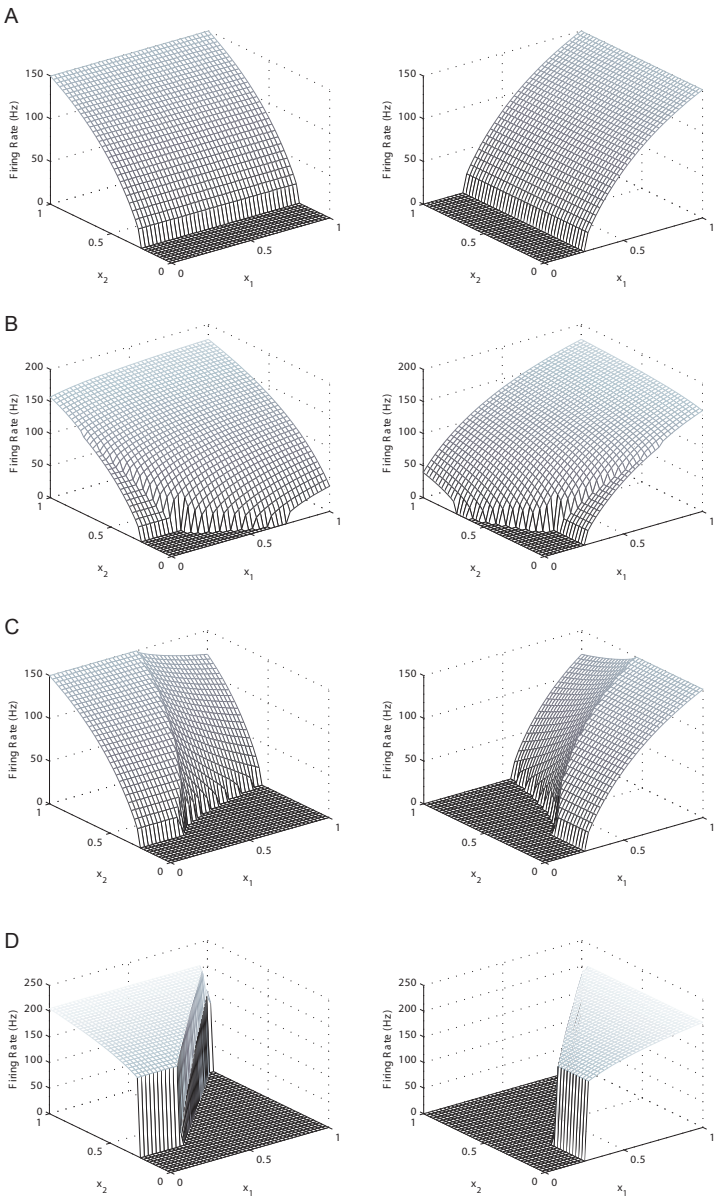

Figure 8.11: Lateral connections modulate feedforward neuronal tuning curves. A, LIF neuron tuning curves with preferred directions x_2 (left) and x_1 (right), and no lateral connections. The remaining panels show how the tuning curves of the same neurons change if they are mutually excitatory (B), mutually inhibitory (C), or if they compete in a winner-take-all manner (D).

rate is maintained close to an equilibrium rate b_{j0}. Specifically,

$$J_j^{bias}(t+1) = J_j^{bias}(t) + \kappa_b(b_{j0} - b_j), \tag{8.7}$$

where κ_b is a learning rate that determines how quickly the bias changes. (It should only change fast enough to maintain the long-term mean firing rate; not so fast that it follows rapid fluctuations in the rate.) The forward weights w_{ji} evolve according to the non-linear Oja rule (see Appendix). However, while in Oja's model the length of each neuron's weight vector is maintained at 1, in the extended model the length is instead maintained at a value γ_j that is different in different neurons. The resulting learning rule for the feedforward weights is

$$w_{ji}(t+1) = w_{ji}(t) + \kappa_w b_j(a_i - \frac{1}{\gamma_j^2} w_{ji}(t) \sum_i w_{ji} a_i), \tag{8.8}$$

where κ_w is a learning rate. For different neurons with parallel weight vectors, γ_j determines the slope of the tuning curve. For a given slope and distribution of inputs, b_{j0} determines the zero-intercept of the tuning curve.

Figure 8.12 shows two neuronal tuning curves at different stages, as they converge on the slopes and intercepts that arise from these intrinsic learning parameters. Notably, this simple learning rule also allows the code to adapt to long-term changes in input statistics, as shown in Figure 8.12B.

This is a very abstract model of homeostasis, but it serves to illustrate the point that diverse neural tuning across a population is compatible with the establishment of neural tuning by synaptic plasticity.

8.6 Discussion

This chapter has shown how Hebbian plasticity can change the preferred directions of neurons within the space they encode. Because synaptic plasticity operates constantly in the brain, one would expect given this result that the tuning of recorded neurons might change during a long experiment, and such changes are indeed observed (e.g. [315]).

Importantly, plasticity can also change the dimension of the coded space. Some of the tasks that the brain performs (e.g. control of movement through space) have a static dimension over an animal's lifetime. However, the ability to re-organize neurons around codes of varying dimensions may be important in networks that develop skilled performance of novel tasks.

This chapter has shown that high-dimensional lateral interactions exert a strong influence on the dimension of the feedforward transform (and consequently the dimension of the post-synaptic code). Many models of Hebbian plasticity employ high-fidelity neurons, and employ lateral inhibition to decorrelate their responses. In population-coding terms, the dimension of the information represented by such

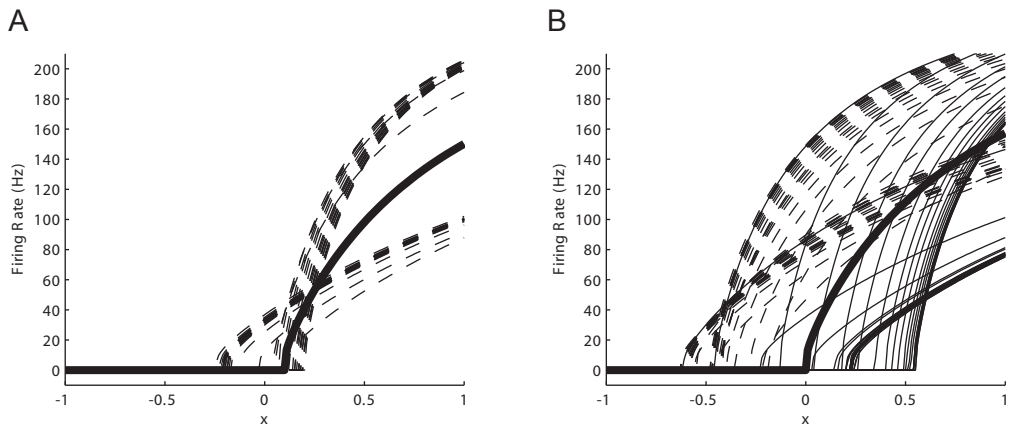

Figure 8.12: Diversity in plasticity parameters leads to diversity in tuning curves. A, The thick line indicates identical tuning of two different LIF neurons before training. Intrinsic parameters in these neurons (discussed in the text) are set so that one converges to a firing threshold of $x = 0.2$, and a maximum firing rate of 200Hz, and the other converges to a firing threshold of x = -0.2 and a maximum firing rate of 100Hz. Dashed lines show progression toward the target tuning over the course of a simulation with uniformly-distributed inputs from x = -1 to 1. B, The homeostatic mechanism allows the tuning curves to adapt to changes in the input statistics. As in (A), the thick solid line indicates the initial tuning curve of both neurons. The dashed lines show progression toward equilibrium tuning with Gaussian-distributed input of $x = -0.5 \pm 0.5$ (SD). Half way through the simulation, the mean changes to 0.5. The thin solid lines show adaptation of the tuning curves to new equilibrium values for this new input distribution.

idealized models is equal to the number of neurons. In contrast, in the absence of lateral interactions, the dimension of the code may be as low as one, or slightly higher, depending on neuronal nonlinearities and input statistics. Sparse lateral inhibition yields codes of intermediate dimension. Self-organizing maps combine lateral inhibition with lateral excitation, and selectively supress coding of noise-like dimensions of the input.

Another key feature of a population code is the diversity of tuning curves among neurons with the same preferred direction. In the context of linear synaptic integration, this diversity determines the computations that can be performed on the encoded information. Section 8.5 introduced a physiologically plausible extension of the Oja rule, in which the values of two learning-rule parameters can be chosen so that a neuron's tuning curve converges to any selected parameterization of the LIF model.

8.6.1 Sparse Coding

An additional parameter of population codes, which was not addressed here, is sparseness of neural activity. In a sparse code, there is little overlap between the tuning curves of different neurons, and few neurons are strongly active at any given time. Olshausen & Field [286] showed that if a network is trained to represent natural images in a sparse manner, the tuning of neurons in the network becomes spatially localized and band-pass. In this sense, the tuning resembles the receptive fields of neurons in the primary visual cortex, suggesting that V1 is optimized for sparse coding of natural scenes. Similar sparse coding is evident in other sensory cortical areas (reviewed by [287]). Decomposing the synaptic weights of the Olshausen & Field [286] model reveals a code with a broad range of singular values, evenly-distributed preferred directions, and decoders that resemble the principal components of natural images (not shown). So it appears that this model narrows the neurons' tuning curves, but essentially retains the dimension of the input signals. Exploration of the relationship between plasticity, sparse coding, and other characteristics of population codes is deferred for future work.

8.6.2 High-Fidelity Neurons as Population Models

In the presence of noise, populations of redundant neurons can represent information with much higher fidelity than individual neurons. One way to model a population of noisy neurons is as a single high-fidelity neuron. This approach is widely used in the relatively complex and long-timescale networks of connectionist cognitive models, in which the inherent computational efficiency is particularly appealing.

However, in such a model, the dimension of represented information is constrained by the number of high-fidelity neurons, and this constraint may influence learning. In other words, it is possible that if the few high-fidelity units in such a

model were replaced with many low-fidelity units, the model would behave differently. The techniques of this chapter could be used to find and correct this type of problem, using the following procedure:

1. Create and train an idealized network of high-fidelity neurons.
2. Convert to a larger network of low-fidelity, spiking neurons using synaptic weights from the idealized network as decoders, and uniformly-distributed encoders.
3. Simulate plasticity in the new network, starting with NEF weights, and determine whether the network remains viable and whether there are any important structural changes.
4. If there are changes, decompose the new weight matrices to discover new decoders, and create a new network using one high-fidelity neuron for each dimension of the synaptic weights.

This would provide a means of testing whether a network of high-fidelity neurons does in fact capture the essential behaviour of a more realistic network of redundant, noisy neurons, and whether the structure of the network changes when artificial constraints on dimensionality are lifted. Furthermore, if different behaviour emerges from the large model, it would then be possible to extract a revised high-fidelity model that captures as much of this behaviour as is desired.

8.6.3 Future Work

It would be interesting to study generalizations of the winner-take-all networks in this chapter. A winner-take-all network is a type of attractor network. The attractors are at certain corners of a hypercube, in which the axes correspond to the firing rates of different neurons – specifically they are along one axis from zero. Networks with attractors at additional (or different) corners can also be constructed. For example, if one post-synaptic neuron excites another one asymmetrically, then the second will always be active if the first one is active, but not *vice versa*. As another example, weaker lateral inhibition would allow multiple units to become active (this is called kWTA, where k is the maximum number of active units). Exploration of effects of Hebbian plasticity on population codes in these more general networks is deferred for future work.

Notably, the Leabra framework [290] combines kWTA activation with both Hebbian and contrastive-Hebbian plasticity. An initial investigation suggests that Leabra networks tend to have high-dimensional weights, even given a task that can be solved with a low-dimensional backpropagation network of the same form (results not shown).

This chapter has focused on unsupervised Hebbian plasticity, which is ubiquitous in the brain. However, supervised learning and reinforcement learning are also important. Clearly, if a sufficiently detailed supervisory signal were available to a network, then the supervisor could shape the code with great flexibility, as illustrated in Figure 8.2. Unlike the Hebbian mechanisms explored here, a supervisory signal could also exert flexible control over the function that is computed by a projection. Similarly, reinforcement signals (which are probably much more common than error signals) provide an alternative basis for modifying synaptic weights based on network performance, and should allow for greater flexibility in shaping decoding vectors.

The lessons learned from the present study of simple plasticity models may help to guide future exploration of more physiologically-detailed models.

8.6.4 Conclusion

Synaptic plasticity is shaped by interactions between cell-intrinsic molecular and spatial factors, and the network properties that shape spiking activity. Much remains to be learned about these mechanisms, and tractable quantitative models are simplistic even in terms of the current state of knowledge. The implications of Hebbian plasticity for population coding will be have to be revisited as these models improve.

Nonetheless, it is encouraging that the simple models of Hebbian plasticity explored here have clear effects on key features of population codes, including their dimension, the distribution of preferred direction vectors, and diversity of the tuning curves.

8.7 Appendix

In Oja's [281] model of Hebbian synaptic plasticity, competition between synapses prevents unbounded growth of the synaptic weights. Weight changes at each time step are normalized as follows:

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\beta[\tilde{\mathbf{w}}(t+1)]}, \tag{8.9}$$

$$\tilde{w}_i(t+1) = w_i(t) + \kappa b a_i, \tag{8.10}$$

where a_i is the firing rate of the i^{th} presynaptic neuron, b is the firing rate of the post-synaptic neuron under consideration, κ is a learning rate, and w_i is the weight of the synapse from the i^{th} presynaptic neuron. The function $\tilde{w}_i(t+1)$ is the synaptic weight that would exist at time $t+1$ in the absence of normalization. Finally, $\beta[\bullet]$ is a function of synaptic weights, which the normalization process keeps constant.

In the commonly-cited form of the Oja rule, $\beta[\tilde{\mathbf{w}}] = ||\tilde{\mathbf{w}}||^2$ is the Euclidian length of the non-normalized weight vector.

Oja showed that for small κ, this weight normalization leads to the update rule,

$$w_i(t+1) = w_i(t) + \kappa \left[a_i b - w_i(t) \left. \frac{\partial \beta[\tilde{\mathbf{w}}(t+1)]}{\partial \kappa} \right|_{\kappa=0} \right]. \tag{8.11}$$

With β defined as above,

$$\begin{aligned} \frac{\partial \beta}{\partial \kappa} =& \frac{\partial}{\partial \kappa} \sqrt{\sum_i [w_i(t) + \kappa a_i b]^2} \\ =& \frac{\sum_i [w_i(t) + \kappa a_i b] a_i b}{\sqrt{\sum_i [w_i(t) + \kappa a_i b]^2}} \\ \left. \frac{\partial \beta}{\partial \kappa} \right|_{\kappa=0} =& \frac{b \sum_i w_i(t) a_i}{\sqrt{\sum_i w_i^2(t)}}. \end{aligned}$$

In the common form of the rule, since the denominator always equals one, and the neurons are linear, this expression simplifies to b^2, yielding

$$w_i(t+1) = w_i(t) + \kappa b(a_i - w_i(t) b). \tag{8.12}$$

In the general case discussed in Section 8.5, with nonlinear neurons and $\beta[\tilde{\mathbf{w}}] = \frac{1}{\gamma}||\tilde{\mathbf{w}}||^2$ (so that $||\tilde{\mathbf{w}}||^2 = \gamma$ at equilibium),

$$w_i(t+1) = w_i(t) + \kappa b(a_i - \frac{1}{\gamma^2} w_i(t) \sum_i w_i(t) a_i). \tag{8.13}$$

With $\gamma = 1$ (as in Section 8.3) this simplifies to

$$w_i(t+1) = w_i(t) + \kappa b(a_i - w_i(t) \sum_i w_i(t) a_i). \tag{8.14}$$

Chapter 9

Conclusions

9.1 Theoretical Principles

This thesis has brought into focus a number of theoretical principles that can potentially improve our understanding of how the basal ganglia work. While models synthesize many concepts into a coherent system, theoretical principles are ideally well-isolated and simple. The principles introduced in the previous chapters are rephrased below as simply as possible.

1. *Compression facilitates selection.* Chapter 2 addressed the important but unresolved question of the role the basal ganglia play in normal brain function. The dominant hypothesis of action-selection was contrasted with reinforcement-driven dimensionality reduction. It was argued that each hypothesis fits more closely with different sets of experimental data, and that the two hypotheses are not mutually exclusive. In particular, a dimensionality reduction network would serve as an effective input stage for a selection network. The dimensionality reduction network would emphasize key contextual information while filtering out information that is likely to be irrelevant, and it would facilitate generalization to novel selection contexts. Whether the basal ganglia actually work in this way is another question that is best addressed experimentally. A means of doing so was proposed.

2. *Computation requires diversity or nonlinear integration.* After Chapter 2, the remainder of the thesis is concerned with the details of computation in basal ganglia networks. The Neural Engineering Framework (NEF) is an appropriate foundation for this work, because it provides a basic set of principles for understanding computation in large networks. However, central to one of these principles is the assertion that computation can be understood in terms of linear decoding. Applying this principal to the basal ganglia is potentially problematic, because many basal ganglia neurons integrate their inputs in a nonlinear manner. One form of nonlinearity, i.e. nonlinear combination of

the input to different major dendritic branches, is strong in the medium spiny neurons of the striatum. Chapter 4 showed that this type of nonlinearity can be understood, in the context of the NEF, as affecting the neuron's tuning curve. Another form of nonlinearity, shunting inhibition, has been considered to play a role in divisive computations. However, Chapter 4 showed that it also supports a flexible form of population coding, which was called the "averaging code". This code is similar to the population codes of the NEF, but it has some advantages. For example, it supports Hebbian learning of computations in very general circumstances, without requiring the inputs to form a tight frame. As shown later (Chapter 7), it can also insulate different network layers from the intrinsic dynamics of other layers. However, the key point of this chapter is that nonlinearities *within* dendritic branches, which apparently present the strongest challenge to the NEF assumption of linear decoding, can actually remove a constraint on linear decoding. Specifically, these nonlinearities allow computation to proceed independently of the diversity of presynaptic tuning curves.

3. *Any diversity will do.* The prevalence of nonlinear input-output relationships within dendritic branches in basal ganglia neurons is not clear (although NMDA receptors, one of the putative mechanisms, are plentiful [202]). Given this uncertainty, the linear-decoding assumption provides a conservative estimate of a network's computational power. Computation *via* linear decoding relies on diversity in the responses of different neurons. However, Chapter 5 shows that this diversity can take on subtle forms. In particular, diverse computations are supported by a population of neurons that fire irregularly at a constant rate, as long as spike-timing correlations between neurons are not too high. In the basal ganglia, these correlations are normally very low, although they increase substantially in Parkinson's disease. The mechanism that produces irregular firing is not critical to this result, but one possible mechanism would be irregularly-shaped tuning curves. This mechanism provides a clear bridge between the results of this chapter and the main body of NEF theory.

4. *Excitation is optional.* The computational power of a neuronal projection is much greater if each presynaptic neuron can effectively excite some of its targets and inhibit others. Previously, Parisien et al. [295] showed that this can be achieved through a combination of excitatory projection neurons and inhibitory interneurons. Chapter 6 shows that the same is true if both the interneurons and the projection neurons are inhibitory. Thus inhibition appears to be essential for maximizing computational flexibility, but excitation does not. The unusual dominance of inhibitory projection neurons in basal ganglia networks may have very minor implications in terms of the computations they support.

5. *Firing dynamics encode history.* Nervous systems are adapted to dynamic environments, and it would be surprising if their computational properties

could be understood in isolation from their dynamic properties. The NEF takes a large step toward integrating these domains. But in doing so, it makes an assumption that does not hold for many basal ganglia neurons, i.e. that neurons' input-output dynamics are dominated by the dynamics of their post-synaptic currents. Contrary to this assumption, the medium spiny neurons of the striatum exhibit dopamine-dependent hysteresis, neurons in the subthalamic nucleus fire in bursts when inhibition is relieved, and neurons in the globus pallidus exhibit varying degrees of adaptation and rebound bursting. Extending the NEF to account for firing dynamics is not straightforward. Chapter 7 shows that the NEF methods apply fairly directly in the simple case of uniform, linear firing-rate adaptation, but that other cases require different approaches. Notably, diverse and nonlinear firing dynamics can be understood as extending the neurons' tuning curves to multiple points in time. This perspective makes it possible to analyse interactions between populations of dynamic neurons much as if they were higher-dimensional static neurons.

6. *Plasticity shapes the code.* The preceding principles apply to the behaviour of mature basal ganglia networks over short time scales. But it is also important to understand how these results relate to changes over longer time scales, as a result of synaptic plasticity. Chapter 8 shows that population-coding concepts, including preferred direction vectors and decoding vectors, have a simple relationship with learned synaptic weights. It is also shown that Hebbian learning rules, in conjunction with lateral connections, can determine both the orientation and the dimension of neurons' preferred direction vectors. These results are particularly relevant for understanding population coding in the striatum, because synaptic weights in the cortico-striatal projection are established by dopamine-gated Hebbian plasticity.

9.2 Future Work

A number of suggestions for future work were outlined in the related chapters. One key suggestion was an experiment for testing whether the striatum represents actions or associated contexts (Chapter 2). Another was the suggestion of expressing dynamic tuning curves in terms of spikes rather than firing rates (Chapter 7). This could be done using spike-triggered principal components of the input (generalizing the widely-used spike-triggered average). Neurons' response functions would then consist of firing probabilities in the input-history space.

In addition to the suggestions made in previous chapters, several further potential avenues of exploration have come to light during this work. These are not directly related to the above principles, but they may help to clarify basal ganglia function in complementary ways. Three of these directions are discussed below.

9.2.1 Compilation of Simultaneous Actions

Motor actions have a hierarchical structure. For example, opening a door is an action, one that might be performed frequently and automatically. However, this action is itself a coordinated collection of movements that could also be considered actions, including termination of gait, reaching, grasping, turning the knob, adjusting balance (in anticipation of force on the hand), and pulling. Coordinating the components of door opening does not require conscious attention. But neither does executing these components individually, or in other well-practiced combinations. Each of these components can also be decomposed further. For example, reaching for the door knob involves flexion of the shoulder, extension of the elbow, and extension of the fingers, each of which can be performed individually.

Interestingly, if an action is decomposed finely enough, the individual elements become more difficult. For example, shoulder flexion involves the coordinated action of several muscles, which are more difficult to activate alone. Specific training is needed [42] to individually activate a single motor unit (i.e. a motor neuron and the associated muscle fibres), despite the fact that skilled coordination of thousands of motor units is effortless.

Common patterns of coordination are automated through experience, beginning prenatally. However, the brain is probably quite limited in its ability to control *novel* combinations of actions. For example, suppose it were possible to describe a good golf swing as a list of ten key components. Even if every one of these components were easy to perform on its own, memorizing the list would be an ineffective way for a novice to prepare for golf.

If the basal ganglia play a role in automating individual actions, they may also automate simultaneous combinations of actions. This role would be consistent with the observation that people with Parkinson's disease are particularly impaired at performing novel, simultaneous combinations of actions [47]. This form of automation is apparently fundamental to complex motor behaviour. If there is a limit to the number of actions that the brain can coordinate in a novel combination, then in order to control a complex movement (e.g. playing a guitar chord), the brain must first automate simpler groups of components. If similar selection processes operate in the prefrontal cortex, this type of automation may also be important for sophisticated cognition.

Limits on the Complexity of Selection

One of the basic unanswered questions in this area is how many actions the brain can execute at once, in a novel combination.[1] An informal pilot experiment was performed as a preliminary investigation of this question. A single subject attempted

[1] Most studies of novel movement coordination have involved only two simultaneous movements. One exception was a study of novel (yet repetitive) simultaneous movements in four limbs [252], which found that these were more successful when the mode of coordination (in-phase or anti-phase) was the same in the upper and lower limbs (see also [227]).

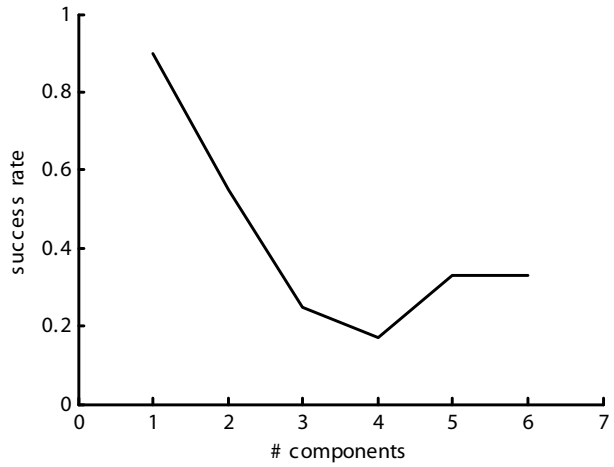

Figure 9.1: Success in simultaneously executing novel combinations of single-joint movements (data from a single subject). The success rate is below 40% for combinations of three or more movements.

to perform simultaneous combinations of up to six low-precision, single-joint movements. The movement combinations were selected in order to be very probably novel to the subject, and so that they did not interfere with each other physically. This subject was unable to reliably execute novel combinations of more than two movements (Figure 9.1). There was occasional success with much larger combinations, up to six movements. In these cases, the subject appeared to mentally rehearse the combination for some time before attempting it, but success was still infrequent.

These preliminary results highlight a sharp limitation on motor selection, or possibly on selection in general. Automaticity, presumably subserved in part by the basal ganglia, clearly mitigates the effect of this limitation on well-practiced movements, which often involve many separable components.

9.2.2 Migration of Procedural Memories

An experiment was suggested in Chapter 2 to determine where the mapping occurs between representations of context and actions. A closely related question is how this mapping is established. A compressed context signal would presumably contain all the information needed to drive automated actions, but the action mapping could not emerge from statistical properties of the context signal alone. The mapping would have to be guided, either by reinforcement or supervision.

Interestingly, during the performance of actions that have yet to be automated, the motor areas of the cortex contain ideal supervisory signals, i.e. commands that drive actions which are adapted to the animal's context. One possibility is that the mapping from context to automated action is established in the thalamo-cortical

projection, and is (at least some of the time) guided by volitional motor commands. Similarly, there are reciprocal connections from the cortex to the thalamic nuclei to which the basal ganglia project. These projections could mediate supervision within the thalamus. Regardless of the site, this type of supervisory mechanism would gradually automate actions that were initially closely attended. This would be consistent with observed migration of activity away from prefrontal areas over the course of learning [322], and re-appearance of prefrontal activity when subjects attend to well-learned movements [189].

This is almost the opposite of another possibility that has been discussed by several authors (e.g. [167, 186]), which is that the output of basal ganglia (assumed to represent actions) supervises connections within the motor cortex, so that selection behaviour that is initially learned in the basal ganglia is ultimately transferred to the cortex. This concept is appealing because the size of the motor cortex (10^9 neurons in humans) probably makes it a better long-term repository of motor memories than the basal ganglia output nuclei (10^5 neurons). The concept is broadly consistent with evidence that motor cortical plasticity is disordered [32], and that cortical motor maps are altered [190] in Parkinson's disease. Primate electrophysiology provides mixed support [138]. Perhaps contrary to the idea, some human imaging studies show increased striatal activity with extended practice of various motor tasks, although further changes can occur after training, so that activity during retention testing is mainly cortical [106]. In adult songbirds, lesions of the basal ganglia do not disrupt song performance, but prevent performance from deteriorating if the bird becomes deaf [105].

Counter-intuitively, these nearly-opposite processes could conceivably operate in parallel. The motor cortex could operate as a constraint satisfaction network, with different inputs conveying commands under cognitive control, contextual information from the basal ganglia, and contextual information from intra-cortical connections [377]. As the cortical network learned, it might reduce its reliance first on cognitive control, and then on basal ganglia output. Alternatively, there might be a more direct chain of supervisory influences, such that prefrontal inputs shape the mapping from compressed context signal to appropriate action, and basal ganglia output shapes a similar mapping from intra-cortical context signals.

These two hypothesized patterns of transfer, and their mechanisms and interactions, constitute another promising subject for future work.

9.2.3 Multi-Scale Modelling

A final direction, in which preliminary progress has been made, relates to modelling across multiple spatial scales.

The NEF unifies a number of domains in systems neuroscience, notably population coding, temporal coding, dynamics, and synaptic plasticity. Some of these relationships were strengthened in previous chapters. But despite its broad scope, the NEF ignores many details at the cellular and molecular levels, while at the same

time it is too complicated for whole-brain modeling. The ideal framework would unify not only systems neuroscience, but all levels of neuroscience, from molecular processes to psychology.

This broader unification would in turn shed further light on systems neuroscience. The behaviour of most neural systems (except perhaps primary sensory and motor areas) is determined largely by interactions with surrounding systems. Similarly, a system's behavior can change dramatically as a consequence of molecular events, which are in turn affected by network activity.

A coherent theory that unifies these multiple scales is a tall order, but computational integration should be straightforward. For example, a simplified whole-brain model could act as a test harness for more detailed models of individual systems, providing realistic inputs, and demanding realistic outputs. Probably the main reason this is not done routinely is because it is inconvenient.

A software system (www.nengo.ca) is being developed (in collaboration with Shu Wu, Terry Stewart, and Chris Eliasmith) to facilitate this process. It provides an implementation of the NEF that can be easily integrated with both higher and lower-level models. The system has recently run a simple hybrid ACT-R/NEF model. The next step is to integrate the NEURON simulation environment, which focuses on single-cell models.

These two points of computational integration should be particularly useful for basal ganglia modelling. As discussed in the introduction, the basal ganglia are central to the ACT-R framework. Furthermore, the NEURON simulator has been used to develop sophisticated models of individual basal ganglia cells (e.g. [256, 388]). The imminent combination of the NEF, a software implementation merged with ACT-R and NEURON, and the new theoretical principles described above, will set the stage for further advances in basal ganglia modelling in the near future.

References

[1] L. F. Abbott. Decoding neural firing and modelling neural networks. *Quart Rev Biophys*, 27:291–331, 1994. 47, 121

[2] L. F. Abbott and P. Dayan. The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11(1):91–101, 1999. 89

[3] L. F. Abbott and S. B. Nelson. Synaptic plasticity: taming the beast. *Nat Neurosci*, 3:1178–1183, 2000. 146, 150

[4] L. F. Abbott and W. G. Regehr. Synaptic computation. *Nature*, 431:796–803, 2004. 124, 143

[5] M. Abeles. Role of the cortical neuron: integrator or coincidence detector? *Isr J Med Sci*, 18:83–92, 1982. 80

[6] M. Abeles, H. Bergman, E. Margalit, and E. Vaadia. Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *J Neurophysiol*, 70(4):1629–1638, 1993. 79

[7] E. Ahissar. Temporal-code to rate-code conversion by neuronal phase-locked loops. *Neural Comp*, 10:597–650, 1998. 80

[8] O. Aizman, H. Brismar, P. Uhlén, E. Zettergren, A. I. Levey, H. Forssberg, P. Greengard, and A. Aperia. Anatomical and physiological evidence for D1 and D2 dopamine receptor colocalization in neostriatal neurons. *Nature Neuroscience*, 3:226–230, 2000. 7

[9] G. Akopian and J. P. Walsh. Reliable long-lasting depression interacts with variable short-term facilitation to determine corticostriatal paired-pulse plasticity in young rats. *J Physiol*, 580:225–40, 2007. 142

[10] J. L. Alberts, M. Saling, C. H. Adler, and G. E. Stelmach. Disruptions in the reach-to-grasp actions of Parkinson's Disease patients. *Exp Brain Res*, 134:353–362, 2000. 4

[11] Roger L. Albin, Anne B. Young, and John B. Penney. The functional anatomy of basal ganglia disorders. *TINS*, 12:366–375, 1989. 5, 6

[12] J. W. Aldridge, R.J. Anderson, and J. T. Murphy. Sensory-motor processing in the caudate nucleus and globus pallidus: a single-unit study in behaving primates. *Can J Physiol Pharmacol*, 58:1192–1201, 1980. 28, 30, 41

[13] G. E. Alexander and M. D. Crutcher. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *TINS*, 13:266–71, 1990. 19

[14] G. E. Alexander and M. D. Crutcher. Neural representations of the target (goal) of visually guided arm movements in three motor areas of the monkey. *J Neurophysiol*, 64(1):164–178, 1990. 28

[15] G. E. Alexander and M. D. Crutcher. Preparation for movement: Neural representations of intended direction in three motor areas of the monkey. *J Neurophysiol*, 64(1):133–150, 1990. 28

[16] G. E. Alexander, M. D. Crutcher, and M. R. DeLong. Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal", and "limbic" functions. volume 85 of *Progress in Brain Research*, pages 119–146. Elsevier Science Publishers, Amsterdam, 1990. 14, 20

[17] G. E. Alexander and M. R. DeLong. Microstimulation of the primate neostriatum. I. physiological properties of striatal microexcitable zones. *J Neurophysiol*, 53(6):1401–1416, 1985. 31

[18] G. E. Alexander and M. R. DeLong. Microstimulation of the primate neostriatum. II. somatotopic organization of striatal microexcitable zones and their relation to neuronal response properties. *J Neurophysiol*, 53(6):1417–1430, 1985. 28, 31

[19] C. Allen and C.F. Stevens. An evaluation of causes for unreliability of synaptic transmission. *PNAS*, 91:10380–83, 1994. 43

[20] D. J. Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge, 1989. 121

[21] C. H. Anderson. Basic elements of biological computational systems. *Int J Modern Physics C*, 5:313–15, 1994. 46

[22] C. H. Anderson and D. C. Van Essen. Neurobiological computational systems. In *IEEE World Congress on Computational Intelligence*, 1994. 46

[23] J. R. Anderson. *How can the human mind occur in the physical universe?* Oxford University Press, 2007. 14, 16

[24] J. R. Anderson, Y. Qin, K.-J. Jung, and C. S. Carter. Information-processing modules and their relative modality specificity. *Cognitive Psychology*, 54:185–217, 2007. 14

[25] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psych Review*, 111:1036–60, 2004. 14

[26] M. E. Anderson and F. B. Horak. Influence of the globus pallidus on arm movements in monkeys. III. timing of movement-related information. *J Neurophysiol*, 54:433–48, 1985. 28

[27] D. Arkadir, G. Morris, E. Vaadia, and H. Bergman. Independent coding of movement direction and reward prediction by single pallidal neurons. *J Neurosci*, 24:10047–56, 2004. 41

[28] N. Aronin, K. Chase, and M. Difiglia. Glutamic acid decarboxylase and enkephalin immunoreactive axon terminals in the rat neostriatum synapse with striatonigral neurons. *Brain Research*, 365:151–8, 1986. 27

[29] H. E. Atallah, D. Lopez-Paniagua, J. W. Rudy, and R. C. O'Reilly. Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nat Neurosci*, 10:126–131, 2007. 38

[30] H. L. Atwood and S. Karunanithi. Diversification of synaptic strength: presynaptic elements. *Nature Reviews Neurosci*, 3:497–516, 2002. 143

[31] B. B. Averbeck, P. E. Latham, and A. Pouget. Neural correlations, population coding and computation. *Nature Reviews Neurosci*, 7:358–66, 2006. 43

[32] S. Bagnato, R. Agostino, N. Modugno, A. Quartarone, and A. Berardelli. Plasticity of the motor cortex in parkinson's disease patients on and off therapy. *Mov Disord*, 21:639–45, 2006. 178

[33] W. Bair and C. Koch. Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Comput*, 8:1185–1202, 1996. 92, 97

[34] B. W. Balleine, M. R. Delgado, and O. Hikosaka. The role of the dorsal striatum in reward and decision-making. *J Neurosci*, 27:8161–65, 2007. 14

[35] I. Bar-Gad and H. Bergman. Stepping out of the box: information processing in the neural networks of the basal ganglia. *Curr Opin Neurobiol*, 11:689–695, 2001. 18

[36] I. Bar-Gad, G. Havazelet-Heimer, J. A. Goldberg, E. Ruppin, and H. Bergman. Reinforcement-driven dimensionality reduction - a model for information processing in the basal ganglia. *J Basic Clin Physiol Pharm*, 11:305–320, 2000. 18

[37] I. Bar-Gad, G. Morris, and H. Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog Neurobiol*, 71:439–473, 2003. 13, 15, 18, 19, 21, 27, 154

[38] M. J. Barber. Information representation in the multi-layer perceptron. In R. F. Bishop, K. A. Gernoth, and N. R. Walet, editors, *150 Years of Quantum Many-Body Theory: A Festschrift in Honour of the 65th Birthdays of John W. Clark, Alpo J. Kallio, Manfred L. Ristig, Sergio Rosati*, pages 319–326. World Scientific, 2001. 147, 148

[39] M. J. Barber, J. W. Clark, and C. H. Anderson. Neural representation of probabilistic information. *Neural Comp*, 15:1843–64, 2003. 46, 48

[40] T. D. Barnes, Y. Kubota, D. Hu, D. Z. Jin, and A. M. Graybiel. Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, 437:1158–1161, 2005. 34

[41] A. G. Barto. Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, chapter 11, pages 215–32. MIT Press, 1995. 25

[42] J. V. Basmajian. Control and training of individual motor units. *Science*, 141:440–41, 1963. 176

[43] P.P. Battaglini, S. Squatrito, C. Galletti, M.G. Maioli, and E. R. Sanseverino. Bilateral projections from the visual cortex to the striatum in the cat. *Exp Brain Res*, 47:28–32, 1982. 23

[44] J. Baufreton, M. Garret, A. Rivera, A. de la Calle, F. Gonon, B. Dufy, B. Bioulac, and A. Taupignon. D5 (not D1) dopamine receptors potentiate burst-firing in neurons of the subthalamic nucleus by modulating an L-type calcium conductance. *J Neurosci*, 23:816–25, 2003. 129

[45] D. G. Beiser, S. E. Hua, and J. C. Houk. Network models of the basal ganglia. *Curr Opin Neurobiol*, 7:185–190, 1997. 14

[46] A. L. Benabid, Z. Ni, S. Chabardes, A. Benazzouz, and P. Pollack. How are we inhibiting functional targets with high frequency stimulation? In K. Kultas-Ilinsky and I. Ilinsky, editors, *Basal Ganglia and Thalamus in Health and Movement Disorders*, pages 309–315. Kluwer Academic/Plenum, New York, 2001. 7

[47] R. Benecke, J. C. Rothwell, J. P. R. Dick, B. L. Day, and C. D. Marsden. Performance of simultaneous movements in patients with Parkinson's disease. *Brain*, 109:739–757, 1986. 4, 176

[48] R. Benecke, J. C. Rothwell, J. P. R. Dick, B. L. Day, and C. D. Marsden. Disturbance of sequential movements in patients with Parkinson's disease. *Brain*, 110:361–379, 1987. 4

[49] A. Benucci, P. F. M. J. Verschure, and P. König. Two-state membrane potential fluctuations driven by weak pairwise correlations. *Neural Computation*, 16:2351–2378, 2004. 83

[50] A. Berardelli, J. Noth, P. D. Thompson, E. L. E. M. Bollen, A. Currà, G. Deuschl, G. van Dijk, R. Töpper, M. Schwarz, and R. A. C. Roos. Pathophysiology of chorea and bradykinesia in Huntington's disease. *Mov Disord*, 14(3):398–403, 1999. 5

[51] H. Bergman, A. Feingold, A. Nini, A. Raz, H. Slovin, M. Abeles, and E. Vaadia. Physiological aspects of information processing in the basal ganglia of normal and parkinsonian primates. *Trends Neurosci*, 21:32–38, 1998. 18, 21

[52] G. S. Berns and T. J. Sejnowski. How the basal ganglia make decisions. In A. R. Damasio et al., editor, *Neurobiology of decision making*. Springer-Verlag, 1996. 14, 20

[53] K. Blatter and W. Schultz. Rewarding properties of visual stimuli. *Exp Brain Res*, 168:541–6, 2006. 30

[54] T. Boraud, E. Bezard, B. Bioulac, and C. E. Gross. Ratio of inhibited-to-activated pallidal neurons decreases dramatically during passive limb movement in the mptp-treated monkey. *J Neurophysiol*, 83:1760–1763, 2000. 21

[55] H. Braak and E. Braak. Pathoanatomy of Parkinson's disease. *J Neurol*, 247 (Supl 2):II/3–II/10, 2000. 32

[56] A. Brand, O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe. Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, 417:543–547, 2002. 80

[57] N. Brenner, W. Bialek, and R. de Ruyter van Steveninck. Adaptive rescaling maximizes information transmission. *Neuron*, 26:695–702, 2000. 143

[58] C. D. Brody and J. J. Hopfield. Simple networks for spike-timing-based computation, with application to olfactory processing. *Neuron*, 37:843–852, 2003. 80

[59] J. W. Brown, D. Bullock, and S. Grossberg. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, 17:471–510, 2004. 23, 108

[60] L. L. Brown, J. S. Schneider, and T. I. Lidsky. Sensory and cognitive functions of the basal ganglia. *Curr Opin Neurobiol*, 7:157–63, 1997. 30

[61] D. V. Buonomano. Decoding temporal information: a model based on short-term synaptic plasticity. *J Neurosci*, 20(3):1129–1141, 2000. 80

[62] G. Burnstock. Cotransmission. *Curr Opin Pharmacology*, 4:47–52, 2004. 103

[63] P. C. Bush and T. J. Sejnowski. Effects of inhibition and dendritic saturation in simulated neocortical pyramidal cells. *J Neurophysiol*, 71:2183, 1994. 66

[64] G. La Camera, A. Rauch, H. R. Lüscher, W. Senn, and S. Fusi. Minimal models of adapted neuronal response in in vivo-like input currents. *Neural Computation*, 16:2101–2124, 2004. 123

[65] M. Carandini and D. J. Heeger. Summation and division by neurons in primate visual cortex. *Science*, 264:1333–6, 1994. 65

[66] C. E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *J Neurosci*, 10:3227–46, 1990. 62

[67] M. Cassidy, P. Mazzone, A. Oliviero, A. Insola, P. Tonali, V. Di Lazzaro, and P. Brown. Movement-related changes in synchronization in the human basal ganglia. *Brain*, 125:1235–1246, 2002. 98

[68] F. S. Chance, L.F. Abbott, and A. D. Reyes. Gain modulation from background synaptic input. *Neuron*, 35:773–82, 2002. 66

[69] J. Chavas and A. Marty. Coexistence of excitatory and inhibitory GABA synapses in the cerebellar interneuron network. *J Neurosci*, 23:2019–31, 2003. 103

[70] C. T. Chen. *Linear System Theory and Design*. Oxford University Press, New York, 1999. 115

[71] C. Chuang, S. Fahn, and S. J. Frucht. The natural history and treatment of acquired hemidystonia: report of 33 cases and review of the literature. *J Neurol Neurosurg Psychiatry*, 72:59–67, 2002. 5

[72] N. Chuhma, H. Zhang, J. Masson, X. Zhuang, D. Sulzer, R. Hen, and S. Rayport. Dopamine neurons mediate a fast excitatory signal via their glutamatergic synapses. *J Neurosci*, 24:972–81, 2004. 23

[73] M. M. Churchland and K. V. Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J Neurophysiol*, 97:4235–57, 2007. 14

[74] M. X. Cohen. Neurocomputational mechanisms of reinforcement-guided learning in humans: A review. *Cognitive, Affective, & Behavioral Neuroscience*, 8:113–25, 2008. 29

[75] M. X. Cohen and C. Ranganath. Reinforcement learning signals predict future decisions. *J Neurosci*, 27:371–8, 2007. 37

[76] G. L. Collingridge, J. T. R. Isaac, and Y. T. Wang. Receptor trafficking and synaptic plasticity. *Nat Rev Neurosci*, 5:952–62, 2004. 146

[77] B. W. Connors and M. J. Gutnick. Intrinsic firing patterns of diverse neocortical neurons. *TINS*, 13:99–104, 1990. 121

[78] L. N. Cooper, N. Intrator, B. S. Blais, and H. Z. Shouval. *Theory of Cortical Plasticity*. World Scientific, 2004. 150

[79] R. Courtemanche, N. Fujii, and A. M. Graybiel. Synchronous, focally modulated β-band oscillations characterize local field potential activity in the striatum of awake behaving monkeys. *J Neurosci*, 23(37):11741–11752, 2003. 98

[80] F. Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *PNAS*, 81:4586–4590, 1984. 98

[81] F. Crick. The recent excitement about neural networks. *Nature*, 337:129–132, 1989. 44

[82] M. D. Crutcher and G. E. Alexander. Movement-related neuronal activity selectively coding either direction or muscle pattern in three motor areas of the monkey. *J Neurophysiol*, 64(1):151–163, 1990. 28

[83] M.D. Crutcher and M.R. DeLong. Single cell studies of the primate putamen II. relations to direction of movement and pattern of muscular activity. *Exp Brain Res*, 53:244–58, 1984. 30, 41

[84] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math Control Signals Systems*, 2:303–14, 1989. 43, 44

[85] U. Czubayko and D. Plenz. Fast synaptic transmission between striatal spiny projection neurons. *PNAS*, 99:15764–69, 2002. 25

[86] S. C. Dakin and P. J. Bex. Role of synchrony in contour binding: some transient doubts sustained. *J Opt Soc Am*, 19(4):678–686, 2002. 80

[87] Y. Dan and M. M. Poo. Spike timing-dependent plasticity of neural circuits. *Neuron*, 44:23–30, 2004. 80

[88] E. D'Angelo, T. Nieus, A. Maffei, S. Armano, P. Rossi, V. Taglietti, A. Fontana, and G. Naldi. Theta-frequency bursting and resonance in cerebellar granule cells: Experimental evidence and modeling of a slow k1-dependent mechanism. *J Neurosci*, 21:759–770, 2001. 120

[89] G. Daoudal and D. Debanne. Long-term plasticity of intrinsic excitability: Learning rules and mechanisms. *Learn Mem*, 10:456–65, 2003. 165

[90] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992. 53

[91] N.D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neurosci*, 8:1704–11, 2005. 16

[92] P. Dayan and L. Abbott. *Theoretical Neuroscience*. MIT Press, 2001. 57, 65, 76, 152, 153

[93] P. Dayan and B. W. Balleine. Reward, motivation, and reinforcement learning. *Neuron*, 36:285–98, 2002. 29

[94] B. C. DeBusk, E. J. DeBruyn, R. K. Snider, J. F. Kabara, and A. B. Bonds. Stimulus-dependent modulation of spike burst length in cat striate cortical cells. *J Neurophysiol*, 78:199–213, 1997. 98

[95] M. R. DeLong. Activity of pallidal neurons during movement. *J Neurophysiol*, 34:414–427, 1971. 28

[96] M. R. DeLong. Primate models of movement disorders of basal ganglia origin. *TINS*, 13(7):281–285, 1990. 5, 6

[97] M. R. DeLong, M. D. Crutcher, and A. P. Georgopoulos. Primate globus pallidus and subthalamic nucleus: Functional organization. *J Neurophys*, 53(2):530–543, 1985. 28

[98] M. R. DeLong and A. P. Georgopoulos. Motor functions of the basal ganglia. Handbook of Physiology Section 1: The Nervous System Volume II: Motor Control, Part 2, chapter 21, pages 1017–1061. 1981. 27, 28

[99] Mahlon R. DeLong and Thomas Wichmann. Circuits and circuit disorders of the basal ganglia. *Arch Neurol*, 64:20–24, 2007. 15

[100] M. Desmurget and R. S. Turner. Testing basal ganglia motor functions through reversible inactivations in the posterior internal globus pallidus. *J Neurophysiol*, 99:1057–76l, 2008. 15, 32

[101] 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*, pages 1–25. MIT Press, Cambridge, 2 edition, 1998. 81

[102] A. Destexhe and D. Paré. Impact of network activity on the integrative properties of neocortical pyramidal neurons in vivo. *J Neurophysiol*, 81:1531–47, 1999. 77

[103] M. Diesmann, M. O. Gewaltig, and A. Aertsen. Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402:529–533, 1999. 101

[104] B. Doiron, A. Longtin, N. Berman, and L. Maler. Subtractive and divisive inhibition: Effect of voltage-dependent inhibitory conductances and noise. *Neural Comp*, 13:227–48, 2000. 66

[105] A. J. Doupe, D. J. Perkel, A. Reiner, and E. A. Stern. Birdbrains could teach basal ganglia research a new song. *TINS*, 28(7):353–363, 2005. 3, 178

[106] J. Doyon, V. Penhune, and L. G. Ungerleider. Distinct contribution of the cortico-striatal and cortico-cerebellar systems to motor skill learning. *Neuropsychologia*, 41:252–62, 2003. 178

[107] B. Draganski, F. Kherif, S. Kloppel, P. A. Cook, D. C. Alexander, G. J. M. Parker, R. Deichmann, J. Ashburner, and R. S. J. Frackowiak. Evidence for segregated and integrative connectivity patterns in the human basal ganglia. *J Neurosci*, 28:7143–52, 2008. 23

[108] B. Dubois and B. Pillon. Cognitive deficits in Parkinson's disease. *J Neurol*, 244:2–8, 1997. 14

[109] R. P. Dum and P. L. Strick. Motor areas in the frontal lobe: the anatomical substrate for the central control of movement. In A. Riehle and E. Vaadia, editors, *Motor Cortex in Voluntary Movements*, chapter 1, pages 3–48. CRC Press, 2005. 14, 31

[110] C. Eliasmith. A unified approach to building and controlling spiking attractor networks. *Neural Comput*, 17:1276–1314, 2005. 52, 121

[111] C. Eliasmith and C. H. Anderson. *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, Cambridge, 2003. 9, 19, 46, 50, 54, 59, 67, 81, 104, 121, 126, 147

[112] S. Fahn. The spectrum of levodopa-induced dyskinesias. *Ann Neurol*, 4(Suppl 1):S2–S11, 2000. 4

[113] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck. Effciency and ambiguity in an adaptive neural code. *Nature*, 412:787–92, 2001. 121, 143

[114] A. Faure, U. Haberland, Françoise Condé, and N. El Massioui. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *J Neurosci*, 25:2771–80, 2005. 37

[115] A. P. Fawcett, J. O. Dostrovsky, A. M. Lozano, and W. D. Hutchison. Eye movement-related responses of neurons in human subthalamic nucleus. *Exp Brain Res*, 162:357–365, 2005. 3, 28

[116] M. B. Feany and W. W. Bender. A Drosophila model of Parkinson's disease. *Nature*, 404:394–398, 2000. 4

[117] G. Fenelon, F. Mahieux, R. Huon, and M. Ziegler. Hallucinations in parkinson's disease: prevalence, phenomenology, and risk factors. *Brain*, 123:733–45, 2000. 31

[118] M. Filion, L. Tremblay, and P. Bédard. Abnormal influences of passive limb movement on the activity of globus pallidus neurons in Parkinsonian monkeys. *Brain Res*, 444:165–176, 1988. 21

[119] J. S. Fitzpatrick, G. Akopian, and J. P. Walsh. Short-term plasticity at inhibitory synapses in rat striatum and its effects on striatal output. *J Neurophysiol*, 85:2088–99, 2001. 142

[120] A. W. Flaherty and A. M. Graybiel. Corticostriatal transformations in the primate somatosensory system. projections from physiologically mapped body-part representations. *J Neurophysiol*, 66:1249–63, 1991. 22

[121] P. Földiák. Adaptive network for optimal linear feature extraction. In *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, 1989. 25, 155

[122] P. Földiák. Forming sparse representations by local anti-hebbian learning. *Biol Cybern*, 64:165–70, 1990. 25

[123] C. A. Fox and J. A. Rafols. The striatal efferents in the globus pallidus and in the substantia nigra. In M. D. Yahr, editor, *The Basal Ganglia*, pages 37–55. Raven Press, New York, 1976. 20

[124] M. J. Frank, B. Loughry, and R. C. O'Reilly. Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, 1:137–60, 2001. 16

[125] M. J. Frank, A. A. Moustafa, H. M. Haughey, T. Curran, and K. E. Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *PNAS*, 104:16311–16, 2007. 26, 33

[126] M. J. Frank, J. Samanta, A. A. Moustafa, and S. J. Sherman. Hold your horses: Impulsivity, deep brain stimulation, and medication in Parkinsonism. *Science*, 318:1309–12, 2007. 27, 36

[127] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly. By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, 306:1940–43, 2004. 33

[128] N. Fuji and A. M. Graybiel. Time-varying covariance of neural activities recorded in striatum and frontal cortex as monkeys perform sequential-saccade tasks. *PNAS*, 102:9032–37, 2005. 28

[129] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J Neurosci*, 2:1527–37, 1982. 14, 42

[130] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science*, 233:1416–1419, 1986. 42, 46, 54

[131] E. Gerardin, S. Lehéricy, J. B. Pochon, S. Tézenas du Montce, J. F. Mangin, F. Poupon, Y. Agid, D. Le Bihan, and C. Marsault. Foot, hand, face and eye representationin the human striatum. *Cereb Cortex*, 13:162–169, 2003. 20

[132] W. Gerstner and W. Kistler. *Spiking Neuron Models. Single Neurons, Populations, Plasticity*. Cambridge University Press, Cambridge, 2002. 129

[133] A. Gillies and G. Arbuthnott. Computational models of the basal ganglia. *Movement Disorders*, 15(5):762–770, 2000. 20

[134] A. J. Gillies and D. J. Willshaw. A massively connected subthalamic nucleus leads to the generation of widespread pulses. *Proc R Soc Lond B*, 265:2101–2109, 1998. 20

[135] C. Günay, J. R. Edgerton, and D. Jaeger. Channel density distributions explain spiking variability in the globus pallidus: A combined physiology and computer simulation database approach. *J Neurosci*, 28:7476–91, 2008. 129

[136] 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(11):1185–1195, 2003. 124

[137] J. Goodall. *Through a window: my thirty years with the chimpanzees of Gombe*. Mariner Books, 1990. 44

[138] A. M. Graybiel. The basal ganglia: learning new tricks and loving it. *Curr Opin Neurobiol*, 15:638–44, 2005. 178

[139] A. M. Graybiel, J. J. Canales, and C. Capper-Loup. Levodopa-induced dyskinesias and dopamine-dependent stereotypies: a new hypothesis. *TINS*, 23(10 Suppl):S71–S77, 2000. 4

[140] A.M. Graybiel. The basal ganglia and chunking of action repertoires. *Neurobiol Learn Mem*, 70:119–36, 1998. 19

[141] A. J. Gruber, S. A. Solla, D. J. Surmeier, and J. C. Houk. Modulation of striatal single units by expected reward: A spiny neuron model displaying dopamine-induced bistability. *J Neurophysiol*, 90:1095–1114, 2003. 137, 138

[142] R. Gütig and H. Sompolinsky. The tempotron: a neuron that learns spike timing-based decisions. *Nat Neurosci*, 9(3):420–428, 2006. 80, 82, 97

[143] K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. I. a new functional anatomy. *Biol Cybern*, 84:401–410, 2001. 16, 17

[144] J. N. Guzmán, A. Hernández, E. Galarraga, D. Tapia, A. Laville, R. Vergara, J. Aceves, and J. Bargas. Dopaminergic modulation of axon collaterals interconnecting spiny neurons of the rat striatum. *J Neurosci*, 23(26):8931–8940, 2003. 21, 25

[145] S. N. Haber, J. L. Fudge, and N. R. McFarland. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J Neurosci*, 20:2369–82, 2000. 23, 104

[146] F. Hadj-Bouziane and D. Boussaoud. Neuronal activity in the monkey striatum during conditional visuomotor learning. *Exp Brain Res*, 153:190–96, 2003. 41

[147] R. H. R. Hahnloser, A. A. Kozhevnikov, and M. S. Fee. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, 419:65–70, 2002. 97

[148] I. Hamada, M. R. DeLong, and N. I. Mano. Activity of identified wrist-related pallidal neurons during step and ramp wrist movements in the monkey. *J Neurophys*, 64(6):1892–1906, 1990. 21, 28

[149] J. E. Hanson and D. Jaeger. Short-term plasticity shapes the response to simulated normal and Parkinsonian input patterns in the globus pallidus. *J Neurosci*, 22:5164–72, 2002. 142

[150] M. Haruno, T. Kuroda, K. Doya, K. Toyama, M. Kimura, K. Samejima, H. Imamizu, and M. Kawato. A neural correlate of reward-based behavioral learning in caudate nucleus: A functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci*, 24:1660–65, 2004. 33

[151] T. Hazy, M. J. Frank, and R. C. O'Reilly. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Phil Trans R Soc B*, 362:1601–13, 2007. 14

[152] D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory.* Wiley, 1949. 146, 150

[153] J. C. Hedreen and M. R. DeLong. Organization of striatopallidal and striatonigral and nigrostriatal projections in the macaque. *J Comp Neurol*, 304:569–595, 1991. 20

[154] S. Hernández-López, J. Bargas, D. J. Surmeier, A. Reyes, and E. Galarraga. D1 receptor activation enhances evoked discharge in neostriatal medium spiny neurons by modulating an L-type Ca2+ conductance. *J Neurosci*, 17(9):3334–3342, 1997. 21

[155] O. Hikosaka, M. Sakamoto, and S. Usui. Functional properties of monkey caudate neurons II. visual and auditory responses. *J Neurophysiol*, 61:799–813, 1989. 30

[156] O. Hikosaka, Y. Takikawa, and R. Kawagoe. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol Rev*, 80(3):953–978, 2000. 3

[157] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Comp*, 18:1527–54, 2006. 46, 121

[158] J. R. Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neurosci*, 1(4):304–309, 1998. 17

[159] G. R. Holt and C. Koch. Shunting inhibition does not have a divisive effect on firing rates. *Neural Comp*, 9:1001–13, 1997. 65

[160] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79:2554–8, 1982. 121

[161] J. J. Hopfield. Pattern recognition computation using action potential timing for stimulus representation. *Nature*, 376:33–36, 1995. 80

[162] F. B. Horak and M. E. Anderson. Influence of globus pallidus on arm movements in monkeys. I. effects of kainic acid-induced lesions. *J Neurophysiol*, 52:290–304, 1984. 32

[163] J. Hore, J. Meyer-Lohmann, and V. B. Brooks. Basal ganglia cooling disables learned arm movements of monkeys in the absence of visual guidance. *Science*, 195:584–6, 1977. 32

[164] J. Hore and T. Vilis. Arm movement performance during reversible basal ganglia lesions in the monkey. *Exp Brain Res*, 39(2):217–228, 1980. 15

[165] J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. MIT Press, Cambridge, 1995. 14, 23, 24, 25, 29

[166] J. C. Houk, C. Bastianen, D. Fansler, A. Fishbach, D. Fraser, P. J. Reber, S. A. Roy, and L. S. Simo. Action selection and refinement in subcortical loops through basal ganglia and cerebellum. *Phil Trans R Soc B*, 2007. 32

[167] J. C. Houk and S. P. Wise. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex*, 2:95–110, 1995. 14, 178

[168] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, 160:106–54, 1962. 42

[169] M. D. Humphries, R. D. Stewart, and K. N. Gurney. A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *J Neurosci*, 26:12921–42, 2006. 16, 20

[170] M. Häusser, G. Major, and G. J. Stuart. Differential shunting of EPSPs by action potentials. *Science*, 291:138–41, 2001. 75

[171] W. D. Hutchison, A. E. Lang, J. O. Dostrovsky, and A. M. Lozano. Pallidal neuronal activity: Implications for models of dystonia. *Ann Neurol*, 53:480–488, 2003. 5

[172] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, and R. Yuste. Synfire chains and cortical songs: temporal modules of cortical activity. *Science*, 304:559–564, 2004. 101

[173] I. A. Ilinsky and K. Kultas-Ilinsky. Neuroanatomical organization and connections of the motor thalamus in primates. In K. Kultas-Ilinsky and I. A. Ilinsky, editors, *Basal Ganglia and Thalamus in Health and Movement Disorders*, pages 77–91. Kluwer Academic/Plenum, New York, 2001. 2, 24

[174] I.A. Ilinsky, H. Yi, and K. Kultas-Ilinsky. Mode of termination of pallidal afferents to the thalamus: A light and electron microscopic study with anterograde tracers and immunocytochemistry in macaca mulatta. *J Comp Neurol*, 386:601–12, 1997. 117

[175] M. Inase, J. A. Buford, and M. E. Anderson. Changes in the control of arm position, movement, and thalamic discharge during local inactivation in the globus pallidus of the monkey. *J Neurophysiol*, 75:1087–1104, 1996. 32

[176] E. M. Izhikevich. Simple model of spiking neurons. *IEEE Trans Neural Networks*, 14:1569–1572, 2003. 82, 84, 85, 121

[177] E. M. Izhikevich. Polychronization: computation with spikes. *Neural Comput*, 18(2):245–282, 2006. 80, 82, 121

[178] E. M. Izhikevich. *Dynamical systems in neuroscience: the geometry of excitability and bursting*. MIT Press, 2007. 121

[179] E. M. Izhikevich, N. S. Desai, E. C. Walcott, and F. C. Hoppensteadt. Bursts as a unit of neural information: selective communication via resonance. *Trends Neurosci*, 26(3):161–167, 2003. 98

[180] E. M. Izhikevich and G. M. Edelman. Large-scale model of mammalian thalamocortical systems. *PNAS*, 105:3593–8, 2008. 121

[181] D. Jaeger, S. Gilman, and J.W. Aldridge. Neuronal activity in the striatum and pallidum of primates related to the execution of externally cued reaching movements. *Brain Research*, 694:111–27, 1995. 19, 28

[182] D. Jaeger, H. Kita, and C. J. Wilson. Surround inhibition among projection neurons is weak or nonexistent in the rat neostriatum. *J Neurophysiol*, 72:2555–8, 1994. 25

[183] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304:78–80, 2004. 143

[184] L. A. Jeffress. A place theory of sound localization. *J Comp Physiol Psychol*, 41:35–39, 1948. 62

[185] D. Joel, Y. Niv, and E. Ruppin. Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15:535–547, 2002. 17, 29

[186] M. S. Jog and Q. Almeida. Basal ganglia: Structure and function. JayPee Brothers Medical Publishers, 2006. 178

[187] R. S. Johansson and I. Birznieks. First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat Neurosci*, 7:170–177, 2004. 80

[188] E. G. Jones. *The Thalamus*. Cambridge, 2007. 24, 104

[189] M. Jueptner and C. Weiller. A review of differences between basal ganglia and cerebellar control of movements as revealed by functional imaging studies. *Brain*, 121:1437–49, 1998. 37, 178

[190] F. A. Kagerer, J. J. Summers, W. D. Byblow, and B. Taylor. Altered corticomotor representation in patients with Parkinson's disease. *Movement Disorders*, 18:919–27, 2003. 178

[191] E. R. Kandel, W. T. Frazier, R. Wazir, and R. E. Coggeshall. Direct and common connection among identified neurons in Aplysia. *J Neurophysiol*, 30:1353–76, 1967. 41

[192] C. Karachi, C. Franc, K. Parain, E. Bardinet, D. Tande, E. Hirsch, and J. Yelnik. Three-dimensional cartography of functional territories in the human striatopallidal complex by using calbindin immunoreactivity. *J Comp Neurol*, 450:122–34, 2002. 22

[193] J. Katayama, N. Akaike, and J. Nabekura. Characterization of pre- and postsynaptic metabotropic glutamate receptor-mediated inhibitory responses in substantia nigra dopamine neurons. *Neuroscience Research*, 45:101–15, 2003. 104

[194] M. Kato and M. Kimura. Effects of reversible blockade of basal ganglia on a voluntary arm movement. *J Neurophysiol*, 68:1516–34, 1992. 32

[195] R. M. Kelly and P. L. Strick. Macro-architecture of the basal ganglia loops with the cerebral cortex: use of rabies virus to reveal multisynaptic circuits. volume 143 of *Progress in Brain Research*, chapter 42, pages 449–60. Elsevier, 2004. 20, 22

[196] A. Kepecs and J. Lisman. Information encoding and computation with spikes and bursts. *Network: Comput Neural Syst*, 14:103–118, 2003. 98, 121

[197] A. Kepecs, M. C. W. van Rossum, S. Song, and J. Tegner. Spike-timing-dependent plasticity: common themes and divergent vistas. *Biol Cybern*, 87:446–458, 2002. 80

[198] A. Kepecs and X.-J. Wang. Analysis of complex bursting in cortical pyramidal neuron models. *Neurocomputing*, 32-33:181–7, 2000. 141

[199] A. A. Kühn, D. Williams, A. Kupsch, P. Limousin, M. Hariz, G. H. Schneider, K. Yarrow, and P. Brown. Event-related beta desynchronization in human subthalamic nucleus correlates with motor performance. *Brain*, 127:735–746, 2004. 98

[200] T. E. Kimber, C. S. Tsai, J. Semmler, B. P. Brophy, and P. D. Thompson. Voluntary movement after pallidotomy in severe Parkinson's disease. *Brain*, 122:895–906, 1999. 15

[201] M. Kimura. Behaviorally contingent property of movement-related activity of the primate putamen. *J Neurophysiol*, 63(6):1277–1296, 1990. 28, 31

[202] H. Kita. Glutamatergic and GABAergic postsynaptic responses of striatal spiny neurons to intrastriatal and cortical stimulation recorded in slice preparations. *Neuroscience*, 70:925–40, 1996. 174

[203] H. Kita and S. T. Kitai. Intracellular study of rat globus pallidus neurons: membrane properties and responses to neostriatal, subthalamic and nigral stimulation. *Brain Res*, 564:296–305, 1991. 129

[204] B. J. Knowlton, J. A. Mangels, and L. R. Squire. A neostriatal habit learning system in humans. *Science*, 273:1399–1402, 1996. 32, 33

[205] P. Knüsel, R. Wyss, P. König, and P. F. M. J. Verschure. Decoding a temporal population code. *Neural Comput*, 16:2079–2100, 2004. 80

[206] S. Kobayashi, R. Kawagoe, Y. Takikawa, M. Koizumi, M. Sakagami, and O. Hikosaka. Functional differences between macaque prefrontal cortex and caudate nucleus during eye movements with and without reward. *Exp Brain Res*, 176:341–55, 2007. 29

[207] C. Koch. *Biophysics of Computation*. Oxford University Press, Oxford, 1999. 48, 88, 94, 123

[208] T. Kohonen. *Self-organizing Maps*. Springer, 2001. 162

[209] B. P. Kolomiets, J. M. Deniau, P. Mailly, A. Ménétrey, J. Glowinski, and A. M. Thierry. Segregation and convergence of information flow through the cortico-subthalamic pathways. *J Neurosci*, 21(15):5764–5772, 2001. 20

[210] T. Koos, J. M. Tepper, and C. J. Wilson. Comparison of IPSCs evoked by spiny and fast-spiking neurons in the neostriatum. *J Neurosci*, 24:7916–22, 2004. 25

[211] M. Krause, W. Fogel, A. Heck, W. Hacke, M. Bonsanto, C. Trenkwalder, and V. Tronnier. Deep brain stimulation for the treatment of Parkinson's disease: subthalamic nucleus versus globus pallidus internus. *J Neurol Neurosurg Psychiatry*, 70:464–470, 2001. 4

[212] S. Y. Kung and K. I. Diamantaras. A neural network learning algorithm for adaptive principal component extraction (APEX). In *International Conference on Acoustics, Speech, and Signal Processing*, 1990. 25, 154, 155, 156

[213] M.-F. Kuo, W. Paulus, and M. A. Nitsche. Boosting focally-induced brain plasticity by dopamine. *Cerebral Cortex*, 2007. 37, 39

[214] C. J. Lacey, J. Boyes, O. Gerlach, L. Chen, P. J. Magill, and J. P. Bolam. GABA-B receptors at glutamatergic synapses in the rat striatum. *Neuroscience*, 136:1083–1095, 2005. 26

[215] B. Lau and P. W. Glimcher. Action and outcome encoding in the primate caudate nucleus. *J Neurosci*, 27:14502–14, 2007. 23, 28

[216] A. D. Lawrence, B. J. Sahakian, and T. W. Robbins. Cognitive functions and corticostriatal circuits: insights from Huntington's disease. *Trends Cog Sci*, 2:379–88, 1998. 14

[217] M. S. Lee, J. O. Rinne, A. Ceballos-Baumann, P. D. Thompson, and C. D. Marsden. Dystonia after head trauma. *Neurology*, 44(8):1374–1378, 1994. 5

[218] R. Legenstein, C. Naeger, and W. Maass. What can a neuron learn with spike-timing-dependent plasticity? *Neural Comp*, 17:2337–2382, 2005. 80

[219] W. Lei, Y. Jiao, N. Del Mar, and A. Reiner. Evidence for differential cortical input to direct pathway versus indirect pathway striatal projection neurons in rats. *J Neurosci*, 24:8289–99, 2004. 26

[220] M. Lengyel and P. Dayan. Hippocampal contributions to control: The third way. In *NIPS*, 2007. 16

[221] N. A. Lesica, C. Weng, J. Jin, C.-I. Yeh, J.-M. Alonso, and G. B. Stanley. Dynamic encoding of natural luminance sequences by LGN bursts. *PLoS Biology*, 4:1201–12, 2006. 121

[222] R. Levy, P. Ashby, W. D. Hutchison, A. E. Lang, A. M. Lozano, and J. O. Dostrovsky. Dependence of subthalamic nucleus oscillations on movement and dopamine in Parkinson's disease. *Brain*, 125:1196–1209, 2002. 98

[223] R. Levy and B. Dubois. Apathy and the functional anatomy of the prefrontal cortex-basal ganglia circuits. *Cereb Cortex*, 16:916–28, 2006. 15

[224] R. Levy, L.-N. Hazrati, M.-T. Herrero, M. Vila, O.-K. Hassani, M. Mouroux, M. Ruberg, H. Asensi, Y. Agid, J. Féger, J. A. Obeso, A. Parent, and E. C. Hirsch. Re-evaluation of the functional anatomy of the basal ganglia in normal and Parkinsonian states. *Neurosci*, 76(2):335–343, 1997. 7

[225] R. Levy, W. D. Hutchison, A. M. Lozano, and J. O. Dostrovsky. High-frequency synchronization of neuronal activity in the subthalamic nucleus of Parkinsonian patients with limb tremor. *J Neurosci*, 20(20):7766–7775, 2000. 79

[226] R. Levy, W. D. Hutchison, A. M. Lozano, and J. O. Dostrovsky. Synchronized neuronal discharge in the basal ganglia of Parkinsonian patients is limited to oscillatory activity. *J Neurosci*, 22:2855–61, 2002. 21

[227] Y. Li, O. Levin, A. Forner-Cordero, and S. P. Swinnen. Interactions between interlimb and intralimb coordination during the performance of bimanual multijoint movements. *Exp Brain Res*, 163:515–26, 2005. 176

[228] O. Lindvall and A. Björklund. Cell therapy in Parkinson's disease. *NeuroRx*, 1:382–393, 2004. 4

[229] J. E. Lisman. Bursts as units of neural information: making unreliable synapses reliable. *TINS*, 20:38–43, 1997. 98

[230] R. Llinás and M. Sugimori. Electrophysiological properties of in vitro Purkinje cell somata in mammalian cerebellar slices. *J Physiol*, 305:171–95, 1980. 120

[231] M. Lévesque and A. Parent. The striatofugal fiber system in primates: A reevaluation of its organization based on single-axon tracing studies. *Proc Nat Acad Sci*, 102(33):11888–11893, 2005. 7

[232] W. Maass, T. Natshläger, and H. Markram. Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput*, 14:2531–2560, 2002. 97, 101, 143

[233] W. A. MacKay. Synchronized neuronal oscillations and their role in motor processes. *Trends Cog Sci*, 1(5):176–183, 1997. 98

[234] K. MacLeod, A. Bäcker, and G. Laurent. Who reads temporal information contained across synchronized and oscillatory spike trains? *Nature*, 395:693–698, 1998. 80

[235] J. C. Magee. Dendritic Ih normalizes temporal summation in hippocampal CA1 neurons. *Nature Neurosci*, 2(6):508–514, 1999. 62, 99

[236] J. C. Magee and E. P. Cook. Somatic EPSP amplitude is independent of synapse location in hippocampal pyramidal neurons. *Nature Neurosci*, 3(9):895–903, 2000. 62, 99

[237] S. Mahon, J.-M. Deniau, and S. Charpier. Relationship between EEG potentials and intracellular activity of striatal and cortico-striatal neurons: an in vivo study under different anesthetics. *Cereb Cortex*, 11:360–73, 2001. 28

[238] S. Mahon, J.-M. Deniau, and S. Charpier. Corticostriatal plasticity: life after the depression. *TINS*, 27:460–67, 2004. 38

[239] S. Mahon, N. Vautrelle, L. Pezard, S. J. Slaght, J.-M. Deniau, G. Chouvet, and S. Charpier. Distinct patterns of striatal medium spiny neuron activity during the natural sleep-wake cycle. *J Neurosci*, 26:12587–95, 2006. 63

[240] L. Maillard, K. Ishii, K. Bushara, D. Waldvogel, A. E. Schulman, and M. Hallett. Mapping the basal ganglia. fMRI evidence for somatotopic representation of face, hand, and foot. *Neurology*, 55:377–383, 2000. 20

[241] E. Marder and A. A. Prinz. Modeling stability in neuron and network function: the role of activity in homeostasis. *BioEssays*, 24:1145–54, 2002. 165

[242] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–15, 1997. 150

[243] H. Markram, Y. Wang, and M. Tsodyks. Differential signaling via the same axon of neocortical pyramidal neurons. *PNAS*, 95:5323–8, 1998. 143

[244] E. Masliah, E. Rockenstein, A. Adame, M. Alford, L. Crews, M. Hashimoto, P. Seubert, M. Lee, J. Goldstein, T. Chilcote, D. Games, and D. Schenk. Effects of α-synuclein immunization in a mouse model of Parkinson's disease. *Neuron*, 46:857–868, 2005. 4

[245] P. B. C. Matthews. The human stretch reflex and the motor cortex. *TINS*, 14:87–91, 1991. 30

[246] J. H. R. Maunsell and D. C. Van Essen. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation. *J Neurophysiol*, 49:1127–47, 1983. 42, 54

[247] C. J. McAdams and J. H. R. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci*, 19:431–41, 1999. 66

[248] 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. 120

[249] N. R. McFarland and S. N. Haber. Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. *J Neurosci*, 20:3798–3813, 2000. 24

[250] N. R. McFarland and S. N. Haber. Thalamic relay nuclei of the basal ganglia form both reciprocal and nonreciprocal cortical connections, linking multiple frontal cortical areas. *J Neurosci*, 22:8117–32, 2002. 24

[251] J. F. Medina, K. S. Garcia, W. L. Nores, N. M. Taylor, and M. D. Mauk. Timing mechanisms in the cerebellum: testing predictions of a large-scale computer simulation. *J Neurosci*, 20(14):5516–5525, 2000. 97

[252] R.L.J. Meesen, N. Wenderoth, and S.P. Swinnen. The role of directional compatibility in assembling coordination patterns involving the upper and lower limb girdles and the head. *Behavioral Brain Research*, 165:262–70, 2005. 176

[253] B. W. Mel. NMDA-based pattern discrimination in a modeled cortical neuron. *Neural Comp*, 4:502–17, 1992. 71

[254] B. W. Mel. Why have dendrites? a computational perspective. In G. Stuart, N. Spruston, and M. Häusser, editors, *Dendrites*, chapter 11, pages 271–89. Oxford University Press, 1999. 63, 71, 72

[255] B. W. Mel. In the brain, the model is the goal. *Nature Neuroscience*, 3:1183, 2000. 9

[256] J. N. Mercer, C. S. Chan, T. Tkatch, J. Held, and D. J. Surmeier. Nav1.6 sodium channels are critical to pacemaking and fast spiking in globus pallidus neurons. *J Neurosci*, 27:13552–13566, 2007. 179

[257] E. Mercuri, J. Atkinson, O. Braddick, S. Anker, F. Cowan, M. Rutherford, J. Pennock, and L. Dubowitz. Basal ganglia damage and impaired visual function in the newborn infant. *Archives of Disease in Childhood*, 77:F111–14, 1997. 31

[258] F. A. Middleton and P. L. Strick. Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. *Science*, 266:458–461, 1994. 14

[259] F. A. Middleton and P. L. Strick. The temporal lobe is a target of output from the basal ganglia. *PNAS*, 93:8683–87, 1996. 31

[260] F. A. Middleton and P. L. Strick. Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews*, 31:236–50, 2000. 22

[261] P. Miller, C. D. Brody, R. Romo, and X. J Wang. A recurrent network model of somatosensory parametric working memory in the prefrontal cortex. *Cerebral Cortex*, 13:1208–1218, 2003. 124

[262] J. W. Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog Neurobiol*, 50:381–425, 1996. 3, 15, 17, 23, 26

[263] J. W. Mink. Basal ganglia dysfunction in Tourette's syndrome: A new hypothesis. *Pediatr Neurol*, 25:190–198, 2001. 5

[264] J. W. Mink. The basal ganglia and involuntary movements. Impaired inhibition of competing motor patterns. *Arch Neurol*, 60:1365–1368, 2003. 15

[265] J. W. Mink and W. T. Thach. Basal ganglia motor control. II. Late pallidal timing relative to movement onset and inconsistent pallidal coding of movement parameters. *J Neurophysiol*, 65:301–29, 1991. 28

[266] J. W. Mink and W. T. Thach. Basal ganglia motor control. III. Pallidal ablation: Normal reaction time, muscle cocontraction, and slow movement. *J Neurophysiol*, 65(2):330–351, 1991. 15, 32

[267] S. J. Mitchell, R. T. Richardson, F. H. Baker, and M. R. DeLong. The primate globus pallidus: neuronal activity related to direction of movement. *Exp Brain Res*, 68:491–505, 1987. 21, 28

[268] A. Münchau and K. P. Bhatia. Pharmacological treatment of Parkinson's disease. *Postgrad Med J*, 76:602–610, 2000. 4

[269] E. Moro, R. J. A. Esselink, J. Xie, M. Hommel, A. L. Benabid, and P. Pollak. The impact on Parksinson's disease of electrical parameter settings in STN stimulation. *Neurology*, 59:706–713, 2002. 4

[270] M. E. Morris, R. Iansek, T. A. Matyas, and J. J. Summers. Stride length regulation in Parkinson's disease: Normalization strategies and underlying mechanisms. *Brain*, 119:551–68, 1996. 32

[271] A. Nambu and R. Llinás. Morphology of globus pallidus neurons: Its correlation with electrophysiology in guinea pig brain slices. *J Comp Neurol*, 377:85–94, 1997. 129

[272] A. Nambu, H. Tokuno, I. Hamada, H. Kita, M. Imanishi, Y. Akazawa, T. Ikeuchi, and N. Hasegawa. Excitatory cortical inputs to pallidal neurons through the cortico-subthalamo-pallidal hyperdirect pathway in the monkey. In A. M. Graybiel, M. R. DeLong, and S. T. Kitai, editors, *The Basal Ganglia VI*, pages 217–223. Kluwer Academic/Plenum, New York, 2002. 3, 27

[273] T. Natschläger and W. Maass. Computing the optimally fitted spike train for a synapse. *Neural Comput*, 13:2477–2494, 2001. 80

[274] T. Natschläger and B. Ruf. Spatial and temporal pattern analysis via spiking neurons. *Network Comput Neural Syst*, 9:319–332, 1998. 80

[275] R. Naud, T. Berger, L. Badel, A. Roth, and W. Gerstner. Quantitative single-neuron modeling: competition 2008. In *COSYNE 2008*, 2008. 48, 121, 123

[276] S. M. Nicola, D. J. Surmeier, and R. C. Malenka. Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu Rev Neurosci*, 23:185–215, 2000. 120

[277] A. Nini, A. Feingold, H. Slovin, and H. Bergman. Neurons in the globus pallidus do not show correlated activity in the normal monkey, but phase-locked oscillations appear in the MPTP model of Parkinsonism. *J Neurphysiol*, 74(4):1800–1805, 1995. 21

[278] E. S. Nisenbaum and T. W. Berger. Functionally distinct subpopulations of striatal neurons are differentially regulated by GABAergic and dopaminergic inputs - I. In vivo analysis. *Neuroscience*, 48:561–78, 1992. 25, 26

[279] E.S. Nisenbaum, T. W. Berger, and A. A. Grace. Presynaptic modulation by GABA-A receptors of glutamatergic excitation and GABAergic inhibition of neostriatal neurons. *J Neurophysiol*, 67:477–81, 1992. 26

[280] E.S. Nisenbaum and C.J. Wilson. Potassium currents responsible for inward and outward rectification in rat neostriatal spiny projection neurons. *J Neurosci*, 15:4449–63, 1995. 19

[281] E. Oja. A simplified neuron model as a principal component analyzer. *J Math Biology*, 15:267–73, 1982. 25, 152, 171

[282] J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34:171–175, 1971. 42

[283] M. S. Okun and J. L. Vitek. Lesion therapy for Parkinson's disease and other movement disorders: Update and controversies. *Mov Disord*, 19(4):375–389, 2004. 7, 15

[284] R. M. Oliveira, J. M. Gurd, P. Nixon, J. C. Marshall, and R. E. Passingham. Micrographia in Parkinson's disease: the effect of providing external cues. *J Neurol Neurosurg Psychiatry*, 63:429–433, 1997. 32

[285] S. J. Olney and D. A. Winter. Predictions of knee and ankle moments of force in walking from EMG and kinematic data. *J Biomech*, 18(1):9–20, 1985. 98

[286] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–9, 1996. 121, 169

[287] B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Curr Opin Neurobiol*, 14:481–7, 2004. 169

[288] L. M. Optican and B. J. Richmond. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. information theoretic analysis. *J Neurophysiol*, 57(1):162–178, 1987. 79

[289] R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz. PVLV: The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, 121:31–49, 2007. 29

[290] R. C. O'Reilly and Y. Munakata. *Computational explorations in cognitive neuroscience*. MIT Press, 2000. 74, 170

[291] A. Parent and F. Cicchetti. The current model of basal ganglia organization under scrutiny. *Mov Disord*, 13:199–202, 1998. 7

[292] A. Parent and L.-N. Hazrati. Anatomical aspects of information processing in primate basal ganglia. *TINS*, 16:111–116, 1993. 16

[293] A. Parent and L. N. Hazrati. Functional anatomy of the basal ganglia. I. The cortico-basal ganglia-thalamo-cortical loop. *Brain Res Rev*, 20:91–127, 1995. 2, 20

[294] A. Parent and L. N. Hazrati. Functional anatomy of the basal ganglia. II. The place of the subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Res Rev*, 20:128–154, 1995. 2, 3, 7, 26

[295] C. M. Parisien, C. H. Anderson, and C. Eliasmith. Solving the problem of negative synaptic weights in cortical models. *Neural Comp*, 20:1473–1494, 2008. 69, 103, 104, 105, 108, 113, 118, 174

[296] V. Pawlak and J. N. D. Kerr. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J Neurosci*, 28:2435–46, 2008. 150

[297] J. P. Pellerin and Y. Lamarre. Local field potential oscillations in primate cerebellar cortex during voluntary movement. *J Neurophysiol*, 78:3502–3507, 1997. 98

[298] G. Percheron and M. Filion. Parallel processing in the basal ganglia: up to a point. *Trends Neurosci*, 14(2):55–56, 1991. 20

[299] D. Plenz. When inhibition goes incognito: feedback interaction between spiny projection neurons in striatal function. *TINS*, 26:436–43, 2003. 104

[300] D. Plenz and S. T. Kitai. Adaptive classification of cortical input to the striatum by competitive learning. In R. Miller and J. R. Wickens, editors, *Brain dynamics and the striatal complex*, chapter 9, pages 165–78. Harwood, 2000. 18, 154

[301] P. Poirazi, T. Brannon, and B. W. Mel. Arithmetic of subthreshold synaptic summation in a model CA1 pyramidal cell. *Neuron*, 37:977–987, 2003. 99

[302] P. Poirazi, T. Brannon, and B. W. Mel. Pyramidal neuron as two-layer neural network. *Neuron*, 37:989–99, 2003. 71

[303] P. Poirazi and B. W. Mel. Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron*, 29:779–96, 2001. 71

[304] R. A. Poldrack, J. Clark, E. J. Paré-Blagoev, D. Shohamy, J. Creso Moyano, C. Myers, and M. A. Gluck. Interactive memory systems in the human brain. *Science*, 414:546–50, 2001. 33

[305] S. A. Prescott and Y. De Koninck. Gain control of firing rate by shunting inhibition: Roles of synaptic noise and dendritic saturation. *PNAS*, 100:2076–81, 2003. 66

[306] A. Raz, E. Vaadia, and H. Bergman. Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine vervet model of Parkinsonism. *J Neurosci*, 20(22):8559–8571, 2000. 21

[307] A. Recchia, P. Debetto, A. Negro, D. Guidolin, S. D. Skaper, and P. Giusti. α-synuclein and Parkinson's disease. *FASEB J*, 18:617–626, 2004. 4

[308] P. Redgrave and K. Gurney. The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neurosci*, 7:967–75, 2006. 30

[309] P. Redgrave, T.J. Prescott, and K. Gurney. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023, 1999. 14, 16

[310] P. Reinagel, D. Godwin, S. M. Sherman, and C. Koch. Encoding of visual information by LGN bursts. *J Neurophysiol*, 81:2558–2569, 1999. 98

[311] P. Reinagel and R. C. Reid. Precise firing events are conserved across neurons. *J Neurosci*, 22(16):6837–6841, 2002. 97

[312] J. N. J. Reynolds and J. R. Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15:507–21, 2002. 17, 33

[313] A. Riehle, S. Grün, M. Diesmann, and A. Aertsen. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, 278:1950–1953, 1997. 79

[314] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: exploring the neural code*. MIT Press, Cambridge, 1997. 47, 79, 131

[315] U. Rokni, A. G. Richardson, E. Bizzi, and H. S. Seung. Motor learning with unstable neural representations. *Neuron*, 54:653–66, 2007. 167

[316] R. Romo and W. Schultz. Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *J Neurophysiol*, 63(3):592–606, 1990. 17

[317] R. Romo and W. Schultz. Role of primate basal ganglia and frontal cortex in the internal generation of movements. III. neuronal activity in the supplementary motor area. *Exp Brian Res*, 91:396–407, 1992. 28

[318] M. Rudolph and A. Destexhe. A fast-conducting, stochastic integrative mode for neocortical neurons in vivo. *J Neurosci*, 23(6):2466–2476, 2003. 62, 99

[319] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by backpropagating errors. *Nature*, 323:533–6, 1986. 44

[320] M. Sahani. A biologically plausible algorithm for reinforcement-shaped representational learning. In *Advances in Neural Information Processing Systems 16*, 2004. 35

[321] M. Sahani and P. Dayan. Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Comp*, 15:2255–79, 2003. 46

[322] K. Sakai, O. Hikosaka, S. Miyauchi, R. Takino, Y. Sasaki, and B. Pütz. Transition of brain activation from frontal to parietal areas in visuomotor sequence learning. *J Neurosci*, 18:1827–40, 1998. 178

[323] E. Salinas and L. F. Abbott. Vector reconstruction from firing rates. *J Comp Neuro*, 1:89–107, 1994. 47

[324] E. Salinas and T. J. Sejnowski. Impact of correlated synaptic input on output firing rate and variability in simple neuronal models. *J Neurosci*, 20(16):6193–6209, 2000. 80

[325] E. Salinas and P. Thier. Gain modulation: A major computational principle of the central nervous system. *Neuron*, 27:15–21, 2000. 66

[326] F. Sato, M. Parent, M. Levesque, and A. Parent. Axonal branching pattern of neurons of the subthalamic nucleus in primates. *J Comp Neurology*, 424:142–152, 2000. 7, 26

[327] E. Schneidman, W. Bialek, and M. J. Berry. Synergy, redundancy, and independence in population codes. *J Neurosci*, 23(37):11539–11553, 2003. 89, 99

[328] W. Schultz. Dopamine neurons and their role in reward mechanisms. *Curr Opin Neurobiol*, 7:191–7, 1997. 23

[329] W. Schultz. Predictive reward signal of dopamine neurons. *J Neurophysiol*, 80:1–27, 1998. 29

[330] W. Schultz. Multiple dopamine functions at different time courses. *Annu Rev Neurosci*, 30:259–88, 2007. 29, 30, 33

[331] W. Schultz, P. Apicella, R. Romo, and E. Scarnati. Context-dependent activity in primate striatum reflecting past and future behavioural events. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, chapter 2, pages 11–27. MIT Press, 1995. 28

[332] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–99, 1997. 14

[333] R. S. Schwab, M. E. Chafetz, and S. Walker. Control of two simultaneous voluntary motor acts in normals and in Parkinsonism. *AMA Arch Neurol Phychiatry*, 72(5):591–598, 1954. 4

[334] J. P. Segundo, G. P. Moore, L. J. Stensaas, and T. H. Bullock. Sensitivity of neurones in Aplysia to temporal pattern of arriving impulses. *J Exp Biol*, 40:643–667, 1963. 80

[335] L. D. Selemon and P. S. Goldman-Rakic. Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *J Neurosci*, 5:776–94, 1985. 22

[336] H. S. Seung. How the brain keeps the eyes still. *Proc Nat Acad Sci*, 93:13339–13344, 1996. 124

[337] M. N. Shadlen and J. A. Movshon. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron*, 24:67–77, 1999. 80

[338] D. Shohamy, C. E. Myers, S. Grossman, J. Sage, M. A. Gluck, and R. A. Poldrack. Cortico-striatal contributions to feedback-based learning: converging data from neuroimaging and neuropsychology. *Brain*, 127:851–9, 2004. 33

[339] A. Sidhu, C. Wersinger, and P. Vernier. Does alpha-synuclein modulate dopaminergic content and tone at the synapse? *FASEB J*, 18:637–647, 2004. 4

[340] W. Singer. Time as coding space? *Curr Opin Neurobiol*, 9:189–194, 1999. 80

[341] R. Singh and C. Eliasmith. Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *J Neurosci*, 26:3667–78, 2006. 121

[342] P. H. Smith. Structural and functional differences distinguish principal from nonprincipal cells in the guinea pig MSO slice. *J Neurophysiol*, 73:1653–67, 1995. 62

[343] Y. Smith, D. Rajua, B. Nandaa, J.-F. Parea A. Galvan, and T. Wichmann. The thalamostriatal systems: Anatomical and functional organization in normal and Parkinsonian states. *Brain Research Bulletin*, 2008. 24

[344] M. A. Sánchez-González, M. A. García-Cabezas, B. Rico, and C. Cavada. The primate thalamus is a key target for brain dopamine. *J Neurosci*, 25:6076–6083, 2005. 37

[345] W. R. Softky and C. Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci*, 13(1):334–350, 1993. 80

[346] P. Somogyi, G. Tamás, R. Lujan, and E. H. Buhl. Salient features of synaptic organisation in the cerebral cortex. *Brain Research Reviews*, 26:113–35, 1998. 104

[347] A. Starr, A. Kang, A. Heath, S. Shimamoto, and S. Turner. Pallidal neuronal discharge in Huntington's disease. In *IBAGS IX*, 2007. 7

[348] P. L. Strick, R. P. Dum, and N. Picard. Macro-organization of the circuits connecting the basal ganglia with the cortical motor areas. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 117–130. MIT Press, Cambridge, 1995. 20, 22

[349] M. Sur and J. L. R. Rubenstein. Patterning and plasticity of the cerebral cortex. *Science*, 310:805–810, 2005. 145

[350] R. Suri. TD models of reward predictive responses in dopamine neurons. *Neural Networks*, 15:523–33, 2002. 23, 24

[351] R. E. Suri and W. Schultz. Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Exp Brain Res*, 121:350–54, 1998. 154

[352] D. J. Surmeier, J. N. Mercer, and C. S. Chan. Autonomous pacemakers in the basal ganglia: who needs excitatory synapses anyway? *Curr Opin Neurobiol*, 15:312–318, 2005. 80, 104

[353] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction*. MIT Press, 1998. 17, 35

[354] C. D. Swinehart and L. F. Abbott. Dimensional reduction for reward-based learning. *Network: Computation in Neural Systems*, 17:235–52, 2006. 36

[355] C. Tang, A. P. Pawlak, V. Prokopenko, and M. O. Wes. Changes in activity of the striatum during formation of a motor habit. *Eur J Neurosci*, 25:1212–27, 2007. 34

[356] J. K. H. Tang, E. Moro, A. M. Lozano, A. E. Lang, W. D. Hutchison, N. Mahant, and J. O. Dostrovsky. Firing rates of pallidal neurons are similar in Huntington's and Parkinson's disease patients. *Exp Brain Res*, 166:230–236, 2005. 7, 80

[357] J. M. Tepper, T. Koós, and C. J. Wilson. GABAergic microcircuits in the neostriatum. *TINS*, 27:662–69, 2004. 25, 104

[358] D. Terman, J. E. Rubin, A. C. Yew, and C. J. Wilson. Activity patterns in a model for the subthalamopallidal network of the basal ganglia. *J Neurosci*, 22(7):2963–2976, 2002. 131, 135

[359] F. Theunissen, J. C. Roddey, S. Stufflebeam, H. Clague, and J. P. Miller. Information theoretic analysis of dynamical encoding by four identified primary sensory interneurons in the cricket cercal system. *J Neurophysiol*, 75:1345–64, 1996. 41

[360] E.L. Thorndike. A proof of the law of effect. *Science*, 77:173–5, 1933. 17

[361] S. Thorpe, A. Delorme, and R. van Rullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715–725, 2001. 80

[362] E. Todorov. On the role of the primary motor cortex in arm movement control. In M. L. Latash and M. F. Levin, editors, *Progress in motor control. Volume Three. Effects of age, disorder, and rehabilitation*, chapter 6, pages 125–166. Human Kinetics, 2004. 14

[363] V. Tomassini, S. Jbabdi, J. C. Klein, T. E. J. Behrens, C. Pozzilli, P. M. Matthews, M. F. S. Rushworth, and H. Johansen-Berg. Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. *J Neurosci*, 27:10259–69, 2007. 16, 24

[364] M. Tomita and J. J. Eggermont. Cross-correlation and joint spectro-temporal receptive field properties in auditory cortex. *J Neurophysiol*, 93:378–392, 2005. 83

[365] J. J. Torres, J.M. Cortes, J. Marro, and H.J. Kappen. Competition between synaptic depression and facilitation in attractor neural networks. *Neural Comp*, 19:2739–55, 2007. 121

[366] L. Tremblay, M. Filion, and P. J. Bédard. Responses of pallidal neurons to striatal stimulation in monkeys with MPTP-induced parkinsonism. *Brain Res*, 498:17–33, 1989. 21

[367] A. Triller and D. Choquet. Surface trafficking of receptors between synaptic and extrasynaptic membranes: and yet they do move! *TINS*, 28:133–39, 2005. 152

[368] B. P. Tripp and C. Eliasmith. Comparison of neural circuits that estimate temporal derivatives. In *Computational and Systems Neuroscience*, 2006. 94

[369] M. J. Tunstall, D. E. Oorschot, A. Kean, and J. R. Wickens. Inhibitory interactions between spiny projection neurons in the rat striatum. *J Neurophysiol*, 88:1263–69, 2002. 25

[370] M. A. Ungless, P. J. Magill, and J. P. Bolam. Uniform inhibition of dopamine neurons in the ventral tegmental area by aversive stimuli. *Science*, 303:2040–42, 2004. 30

[371] N. Urbain, D. Gervasoni, F. Soulière, L. Lobo, N. Rentéro, F. Windels, B. Astier, M. Savasta, P. Fort, B. Renaud, P. H Luppi, and G. Chouvet. Unrelated course of subthalamic nucleus and globus pallidus neuronal activities across vigilance states in the rat. *Eur J Neurosci*, 12:3361–3374, 2000. 7

[372] E. Vaadia, I. Haalman, M. Abeles, H. Bergman, Y. Prut, H. Slovin, and A. Aertsen. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature*, 373:515–18, 1995. 18

[373] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz. Cortical control of a prosthetic arm for self-feeding. *Nature*, 453:1098–1101, 2008. 42

[374] L. Vercueil, P. Pollak, V. Fraix, E. Caputo, E. Moro, A. Benazzouz, J. Xie, A. Koudsie, and A. L Benabid. Deep brain stimulation in the treatment of severe dystonia. *J Neurol*, 248:695–700, 2001. 5

[375] J. L. Vitek, V. Chockkan, J. Y Zhang, Y. Kaneoke, M. Evatt, M. R. DeLong, S. Triche, K. Mewes, T. Hashimoto, and R. A. E. Bakay. Neuronal activity in the basal ganglia in patients with generalized dystonia and hemiballismus. *Ann Neurol*, 46:22–35, 1999. 5

[376] S. Wagner, M. Castel, H. Gainer, and Y. Yarom. GABA in the mammalian suprachiasmatic nucleus and its role in diurnal rhythmicity. *Nature*, 387:598–603, 1997. 103

[377] C. Weber, S. Wermter, and M. Elshaw. A hybrid generative and predictive model of the motor cortex. *Neural Networks*, 19:339–53, 2006. 178

[378] K. K. Wenger, K. L. Musch, and J. W. Mink. Impaired reaching and grasping after focal inactivation of globus pallidus pars interna in the monkey. *J Neurophysiol*, 82:2049–60, 1999. 32

[379] M. B. Westover, C. Eliasmith, and C. H. Anderson. Linearly decodable functions from neural population codes. *Neurocomputing*, 44-46:691–6, 2002. 57

[380] T. G. Weyand, M. Boudreaux, and W. Guido. Burst and tonic response modes in thalamic neurons during sleep and wakefulness. *J Neurophysiol*, 85:1107–18, 2001. 120

[381] A. L. Whone, R. Y. Moore, P. P. Piccini, and D. J. Brooks. Plasticity of the nigropallidal pathway in Parkinson's disease. *Ann Neurol*, 53:206–13, 2003. 36

[382] T. Wichmann, H. Bergman, and M. R. DeLong. The primate subthalamic nucleus. I. Functional properties in intact animals. *J Neurophysiol*, 72(2):494–506, 1994. 28

[383] T. Wichmann and M. R. DeLong. Physiology of the basal ganglia and pathophysiology of movement disorders of basal ganglia origin. In R. L. Watts and W. C. Koller, editors, *Movement Disorders. Neurologic Principles & Practice*, pages 101–112. McGraw-Hill, New York, 2 edition, 2004. 5, 21

[384] J. R. Wickens and D. E. Oorschot. Neural dynamics and surround inhibition in the neostriatum: a possible connection. In R. Miller and J. R. Wickens, editors, *Brain dynamics and the striatal complex*, chapter 7, pages 141–50. Harwood, 2000. 25

[385] S. R. Williams and G. J. Stuart. Site independence of EPSP time course is mediated by dendritic Ih in neocortical pyramidal neurons. *J Neurophysiol*, 83:3177–3182, 2000. 62, 99

[386] S. R. Williams and G. J. Stuart. Dependence of EPSP efficacy on synapse location in neocortical pyramidal neurons. *Science*, 295:1907–10, 2002. 69

[387] C. J. Wilson. Basal ganglia. In G. M. Shepherd, editor, *The Synaptic Organization of the Brain*, pages 361–414. Oxford University Press, Oxford, 5 edition, 2004. 2, 18, 24

[388] J. A. Wolf, J. T. Moyer, and L. H. Finkel. Dopaminergic modulation and afferent input integration in a computational model of the nucleus accumbens medium spiny neuron. In *IBAGS IX*, 2007. 63, 179

[389] B. D. Wright, K. Sen, W. Bialek, and A. J. Doupe. Spike timing and the coding of naturalistic sounds in a central auditory area of songbirds. *arXiv:physics*, page 0201027, 2002. 79

[390] X. Xie and H. S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural Comp*, 15:441–454, 2003. 44

[391] B. Yang, J. D. Slonimsky, and S. J. Birren. A rapid switch in sympathetic neurotransmitter release properties mediated by the p75 receptor. *Nat Neurosci*, 5:539–45, 2002. 104

[392] J. Yelnik, C. François, G. Percheron, and D. Tandé. A spatial and quantitative study of the striatopallidal connection in the monkey. *Neuroreport*, 7:985–8, 1996. 20

[393] H. H. Yin, M. I. Davis, J. A. Ronesi, and D. M. Lovinger. The role of protein synthesis in striatal long-term depression. *J Neurosci*, 26:11811–20, 2006. 38

[394] H. H. Yin and B. J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7:464–76, 2006. 38

[395] H. H. Yin, B. J. Knowlton, and B. W. Balleine. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur J Neurosci*, 19:181–9, 2004. 16, 34, 37

[396] H. H. Yin, B. J. Knowlton, and B. W. Balleine. Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behavioral Brain Research*, 166:189–96, 2006. 34

[397] H. H. Yin, S. B. Ostlund, B. J. Knowlton, and B. W. Balleine. The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22:513–23, 2005. 16

[398] T. C. T. Yin and J. C. M. Chan. Interaural time sensitivity in medial superior olive of cat. *J Neurophysiol*, 64(2):465–488, 1990. 62, 80

[399] K. Yoshida, D. Watanabe, H. Ishikane, M. Tachibana, I. Pastan, and S. Nakanishi. A key role of starburst amacrine cells in originating retinal direction selectivity and optokinetic eye movement. *Neuron*, 30:771–780, 2001. 103

[400] K. K. L. Yung, A. D. Smith, A. I. Levey, and J. P. Bolam. Synaptic connections between spiny neurons of the direct and indirect pathways in the neostriatum of the rat: Evidence from dopamine receptor and neuropeptide immunostaining. *Eur J Neurosci*, 8:861–9, 1996. 27

[401] R. S. Zemel, P. Dayan, and A. Pouget. Probabilistic interpretation of population codes. *Neural Comp*, 10:403–430, 1998. 46

[402] K. Zhang, I. Ginzburg, B. L. McNaughton, and T. J. Sejnowski. Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *J Neurophysiol*, 79:1017–44, 1998. 61

[403] T. Zheng and C. J. Wilson. Corticostriatal combinatorics: The implications of corticostriatal axonal arborizations. *J Neurophysiol*, 87:1007–17, 2002. 18, 22

[404] E. Zohary, M. N. Shadlen, and W.T. Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370:140–43, 1994. 43

[405] R. S. Zucker and W. G. Regehr. Short-term synaptic plasticity. *Annu Rev Physiol*, 64:355–405, 2002. 124