Eye Position Modulation of Visual Cortex and the Sensory Set hypothesis

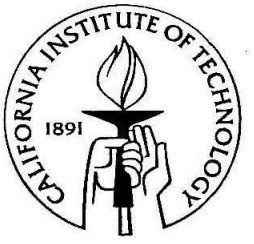
Thesis by

David Rosenbluth

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2002

(Submitted June 21, 2001)

Acknowledgements

I would like to express my gratitude firstly to my thesis advisor John Allman, who supported me over the years of my research both intellectually and financially. I admire his combination of forthrightness and creativity even more now at the completion of my thesis than I did when I first came to his lab. I would also like to thank Yaser Abu-Mostafa for giving me encouragement and more importantly for spending time with me which he could scarcely afford. I thank my other committee members, Mark Konishi, Pietro Perona, and Gilles Laurent for their helpful comments and discussions over the years. Their research and teaching has served as a model for me. Joe Bogen, for whom I served as teaching assistant for several years, has counseled me in both personal and professional matters. He has helped me put things in both philosophical and historical perspective. I owe my thanks to Richard Jeo who graduated from our lab, and to whom I was apprentice. I learned a great deal from him about electrophysiology and how to survive as a graduate student. This type of work is necessarily a group effort and so it would not have been possible without the help of Hollie Weld who kept the animals healthy and happy, contributed to the data collection, and dealt with many things that no one else wanted to deal with, including mischievous monkeys and unreasonable graduate students. I learned a great deal from her. Atiya Hakeem also contibuted to the data collection and in the care of the animals. I thank Romi Nijhawan and Shin Shimojo for the use of the positioning device used in these experiments and Terry Sejnowski, Emilio Salinas, Javier Movellan for helpful discussions.

My wife Melissa, and my two daughters, Imogen and Phoebe, deserve my special thanks for having confidence in me, for being stoic, and for providing happiness in my daily life. We owe our thanks to the ARCS foundation for providing financial support to our family without which this thesis would not be possible.

Abstract

What we see depends on where we look. This is obvious as a statement about the nonuniformity of our external visual environment. But it is also true, in a much less obvious sense, as a statement about the internal neurophysiology of the visual system. What we see depends on where we look in the neurophysiological sense that eye position signals have a dramatic effect on the responsiveness of visual cortical neurons. This thesis empirically studies the way in which point of regard (what point in space the eyes are fixating) influences neurons in visual cortical areas V1 and V4 and then presents a theoretical exploration of how these two different ways in which *"What we see depends on where we look"* might be functionally intertwined.

The empirical data presented here adds to the growing body of evidence that eye position signals are ubiquitous in visual cortex, an observation which reopens speculation about the functional role that these signals might play in different visual cortical areas. The presence of eye position signals in visual areas of the ventral visual processing stream raises the possibility that these signals might facilitate object identity. Eye position signals might be exploited by visual cortex as a conditioned stimulus, which can become functionally linked to the responses of visual cortical neurons (unconditional response) through repeated pairing with the unconditioned stimulus, the retinal stimulus, in a classical conditioning paradigm. In this way the visual system would be capable of learning systematic relationships between point of regard and statistical characteristics of the visual environment. The learned response to the conditioned stimulus could then be exploited as a preparatory signal, to speed or otherwise alter visual processing to suit the current context. In exploring this theoretical viewpoint, we discuss the circumstances under which context dependent coding provides advantages and how a code switching strategy might be implemented through physiological parcellation mediated by gain control. Eye position signals are here considered to be one among many different types of extra-retinal signals present in visual cortical areas, whose presence might be similarly exploited. As such, the data and theory presented here should be considered as contributing to the broader literature on the influence of signals from outside the classical receptive field.

Contents

List of Figures

Chapter 1 Influence of Point of Regard on Neurons in Ventral Visual Cortex

Perception may be regarded as primarily the modification of anticipation.

Art and Illusion, E.H. Gombrich

The eye of a master does more work than both his hands.

Benjamin Franklin

1.1 Introduction

Vision in primates is an active process in which visual information is sought out in the environment and exploited in a purposeful and opportunistic way.[1] The evolution of active vision in primates involved a functionally linked set of changes: foveated retinas; frontal placement of eyes; enlargement and greater parcellation of visual cortical areas; and specialization of cortical areas for visual guidance of muscle movement. These changes gave rise to the primate capacities for high acuity frontal vision and eye-hand coordination, endowing early primates with many advantages in their ecological role as visual predators in the fine branch niche[7]. The primate ability to learn novel visually guided behaviors requires learning not only specialized motor patterns directly related to the behavior, but also specialized eye movement patterns suited to the guidance of the behavior, and the efficient extraction of the particular visual features needed for the execution of the behavior[49]. While the term "active vision" usually refers to the role that eye movements play in the selection of information entering the visual system, the processing of this information in visual cortex may be as active as the eye itself. The activity of both the eye and the processing of the information conveyed by the eye may be actively coordinated. Active changes to the processing of visual information may constitute a "covert" component of active vision.

[1] Active sensation, as a strategy for acquiring information about the environment, has evolved in many sensory modalities and in many organisms. For example, many mammals exploit active movement of the pinnae in audition, and active movement of whiskers in tactile perception. This strategy is used with compact sensor arrays that are densely populated with receptor elements, well suited to the gathering information from a localized region of the environment with very high acuity, but requiring a system for organizing the movement of the sensor in order to sample the environment effectively.

Signals from a variety of sources indicating eye position, threat, and reward, are present in visual cortex[9; 21; 60; 39]. These signals often precede and are indicative of a change in behavioral or sensory context[56]. Eye movements, for instance, typically precede motor actions of visually guided behaviors by about a second, making the eye position signal an important predictor of upcoming visual information entering the processing stream[49]. As a result of learning mechanisms present in cortex, such predictive signals, would tend to influence neuronal responses[72; 5]. The visual system has the capacity to adapt in an eye position dependent manner, a capacity that plays an important role in learning specialized visual tasks. Psychophysical phenomena in which "a new perception is conditioned to the eye position stimulus" have been described as **situational or conditioned aftereffects**[47]. It is plausible that a form of conditioned learning of associations between

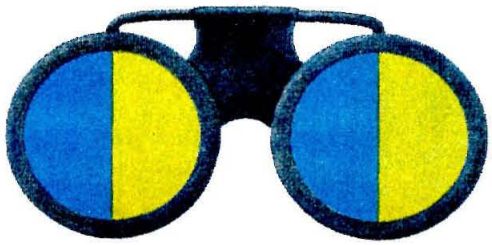


Figure 1.1: Kohler's Goggles
Colored Goggles devised by Kohler create a blue-tinted world when the wearer looks to the left and a yellow-tinted world when he looks to the right. If the goggles are worn for several weeks, the eye adapts and the color distortions tend to disappear. Somehow the visual system learns to introduce the proper correction according to whether the eyes are turned to the left or right. Figure from Experiments with Goggles, Ivo Kohler, Scientific American, 1962

eye position and visual stimulus characteristics is the mechanism underlying these effects, but the anatomical locus and physiological basis for these phenomena remains little explored. This thesis explores the influence of extra-retinal eye position signals, indicating the point in three-dimensional space at which the eyes are directed, on the responsiveness of neurons in V1, V2, and V4.

The presence of eye position signals in visual cortex has been known since the 70's. Profound spatial deficits found in clinical cases of damage to posterior parietal cortex motivated the search for and discovery of neurons which were both responsive to visual stimuli and

influenced by eye position information[10; 70]. The success of this line of research and the associated coordinate transformation theory has influenced where subsequent research has looked for this phenomena and how its functional relevance has been interpreted. In the interval since these seminal studies, further research has found similar eye position modulation of neurons in earlier areas along the dorsal visual processing pathway, and more recently studies have extended these findings to the earliest stages of cortical and sub cortical visual processing[16; 48; 84; 78]. Studies of the modulatory effect of distance cues on cells in both V1 and V4, have indicated that extra-retinal signals related to vergence and accommodation are also present in areas along the ventral visual pathway.[23]. The current study explores the influence of all three spatial parameters, horizontal, vertical, and depth eye position signals, and their interactions. The data presented here deepens the understanding of the influence of these signals by comparing and contrasting the effects found in different visual cortical areas, and contributes to the growing body of evidence that eye position signals are ubiquitous in visual cortex. The difference between the distribution of modulation found in different areas, the fact that this type of modulation is prominent in the ventral pathway, and the lack of clinical evidence which would support a role for these eye position signals in the ventral stream similar to that proposed for the dorsal stream, leads us to propose an alternative role for these signals in ventral visual processing.

Anatomical data suggests that the modulatory eye position signals observed in these experiments may originate in frontal areas commanding saccadic eye position and movement. The convergent termination pattern of eye position inputs, dopamenergic inputs (implicated in reward conditioning), and amygdalar inputs (implicated in aversive conditioning), in layers 5 and 6 of visual cortical areas V2 and V4, provides a potential substrate for learning of correlations between eye position and the nature of visual information entering the visual processing stream.

1.2 Anatomical and Physiological Background

1.2.1 Oculomotor System

While most animals have gaze stabilization[2] mechanisms which align the retina with targets in the external world, gaze shifting[3] mechanisms typically are found only in vertebrates with retinal sub regions specialized for higher acuity, such as the primate foveated retina. In this section we focus on the anatomy of the saccade and gaze-holding circuitry.

In primates, eye movements are controlled via six extra ocular muscles arranged in three antagonistic pairs:

- The medial and lateral rectus muscles controlling the horizontal position of each eye.
- The superior and inferior rectus muscles controlling vertical position of each eye.
- The superior and inferior oblique muscles controlling rotation of the eyes about the line of sight.

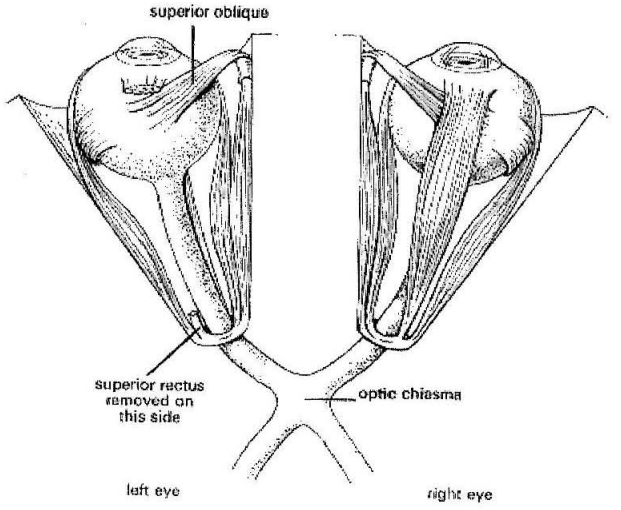


Figure 1.2: Diagram of Eye Muscles

Diagram showing the six extra ocular muscles controlling eye movements. Taken from Eye and Brain by R.L. Gregory[33].

[2]Vestibulo-ocular system and optokinetic system

[3]Version consisting of Saccadic and Smooth pursuit systems, and vergence

In their relaxed state, the passive elastic properties of the muscles create spring-like forces which draw the eye into a central position. Moving the eye from this central position requires overcoming the resistance of the orbit to motion and accelerating the eye. Maintaining the eye in a position, other than the central position, requires a static force to counteract the spring-like tension of the muscles. Eye position is a linear function of the firing rates of oculomotor motor neurons while the eye is stationary, with each motor neuron having a characteristic eye position at which it begins to fire. High frequency bursts of activity generate the dynamic force produced during an eye movement, while tonic activity maintains eye position.

The medial, superior, and inferior recti are innervated by the oculomoter nerve (cranial nerve III); the superior oblique is innervated by the troclear nerve (cranial nerve IV); and the lateral rectus is innervated by the abducens nerve (cranial nerve VI). The third, fourth, and sixth cranial nuclei containing these motor neuron somata are interconnected by a pathway called the medial longitudinal fasciculus (MLF), enabling unilateral coordination of extra ocular muscle activity. In vertical movements of the eyes, due to a slight torsion generated by the superior and inferior rectus muscles, there is a need for coordination not only between the muscles within this antagonistic pair, but also between this pair and the oblique pair which produces a corrective torsional force. In horizontal version movements of the eyes, there is coordination not only unilaterally between the medial and lateral rectus antagonistic pair of muscles of an eye, but also bilaterally between the medial rectus of one eye and the lateral rectus of the other eye. Vergence eye movements also result in a different sort of bilateral coordination in which the medial and lateral rectii of the two eyes operate in concert.

Premotor burst neurons in the paramedian pontine reticular formation (PPRF) activate eye muscle motoneurons during horizontal saccades. Vertical saccades are controlled by premotor burst neurons of the rostral interstitial nucleus of the medial longitudinal fasciculus (MLF). Lesions to the prepositus nucleus of the hypoglossal nerve results in a condition where the eye drifts back to its central position after making saccades, suggesting that the tonic activity of the neurons of this nucleus are responsible for generating the static force signals necessary for maintaining eye position. Integration of the different incoming velocity signals to generate horizontal eye position information involves the coordination of the nuclei prepositi hypoglossi and the medial vestibular nuclei on both sides of the

brain stem as well as the cerebellar flocculus. The interstitial nucleus of Cajal is thought to provide the velocity to position integration for vertical eye movements. Neurons of the rostral midbrain reticular formation form a center for vergence control. Vergence burst-tonic neurons of this area are good candidates for integrators of vergence velocity signals.

The existence of separate systems controlling vergence and version postulated by Ewald Herring, whose outputs are combined during gaze shifts to produce the final binocular motor command[38], are well supported by lesion studies in which deficits to either vergence or version can be produced. A population of neurons in the oculomotor nucleus discharge with vergence, accommodation, or both. This type of behavior reflects the fact that accommodation and vergence are not independent processes. Vergence is altered when accommodation is altered, even in monocular viewing (accommodative-vergence). Likewise, accommodation is altered when vergence is altered (vergence-accommodation). But there is a great deal of evidence that vergence response generally matches vergence demand closely (i.e., fixation disparity is very small), while accommodative response matches accommodative demand less so. Hence quantitatively the influence of disparity on vergence and perhaps also on accommodation is stronger than that of blur[44]. The oculomotor system undergoes three changes as fixation distance decreases, which are collectively known as the near response.

- Vergence: the two eyes converge on the fixation point to minimize binocular disparity.
- Accommodation: the lens of the eye accommodates to minimize blur on the retina.
- Constriction: the pupils transiently constrict causing increase in depth of field.

The involvement of pupillary constriction in the near response may be due to involvement of the pretectal olivary nucleus/nucleus of the optic tract complex.

The pontine and mesencephalic circuits providing the motor signals for saccades are themselves driven by inputs from superior colliculus and also receive direct cortical input from the Frontal Eye Field. Neurons of the intermediate layer of the most rostral portion of the superior colliculus discharge tonically during active visual fixation. These neurons project to caudal parts of the colliculus and to the dorsal raphe nucleus where they inhibit saccade generation. Activity of neurons in the intermediate layer of caudal superior colliculus precede eye movements. These neurons are arranged in a *map* of potential eye movements, and focal electrical stimulation within this map evokes saccades into the movement fields of the stimulated neurons. Parameters of eye movements appear to be encoded

by the activity of populations of neurons within this map. Neurons of intermediate layers

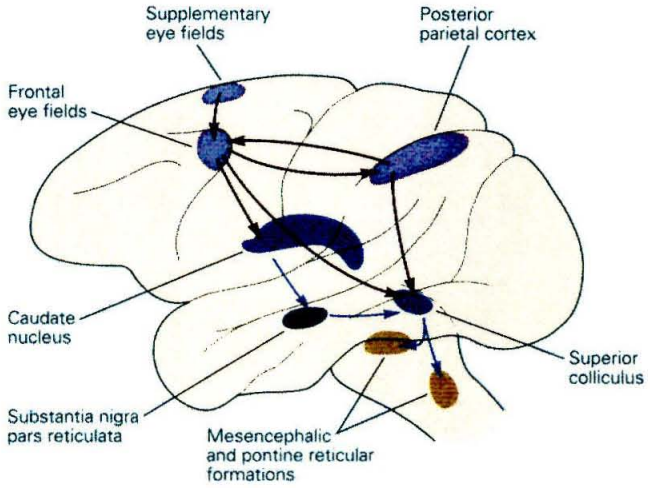


Figure 1.3: Anatomy of Saccadic Eye Movements
Schematic diagram showing the main pathways governing the cortical control of saccades. Taken from Kandel and Schwartz

of superior colliculus receive direct cortical input from both LIP in parietal cortex and the Frontal Eye Fields in frontal cortex. The inputs from LIP are thought to play an important role in linking visual attention with saccade behavior. Frontal Eye Fields play an important role not only in saccades but also in the control of gaze-fixation, via their projection onto omnipause neurons in the nucleus raphe interpositus[18]. Visual neurons in FEF respond vigorously to stimuli that will be targets of saccades; Movement related neurons of the Frontal Eye Field fire only before saccades that are relevant to the monkeys behavior; and visuomovement neurons of FEF discharge most before visually guided saccades. FEF influences superior colliculus both directly, through projections to intermediate layers, and indirectly through projections to caudate nucleus which results in a release of superior colliculus from inhibition by the substantia nigra. Lesions to superior colliculus produces only transient damage to the saccadic system due to the presence of the direct projection from FEF to the brain stem.

Oculomotor cues, such as accommodation and vergence, can be quite effective signals for determining distance at close ranges. Psychophysical studies have demonstrated that subjects are able to make accurate distance judgments in the absence of pictorial cues for distances of less than one meter[36]. At larger distances pictorial cues become increasingly

important in the judgment of distance. This may be due in part to the fact that the magnitude of the accomodative power and the vergence angle declines exponentially with respect to distance. For example, the difference in lens power for shifting fixation from 20cm (the approximate near point for an adult human observer) to one meter is 5 diopters. On the other hand, shifting fixation from 1 meter to 5 meters results in a change of only 0.8 diopters as shown graphically in figure 1.4.

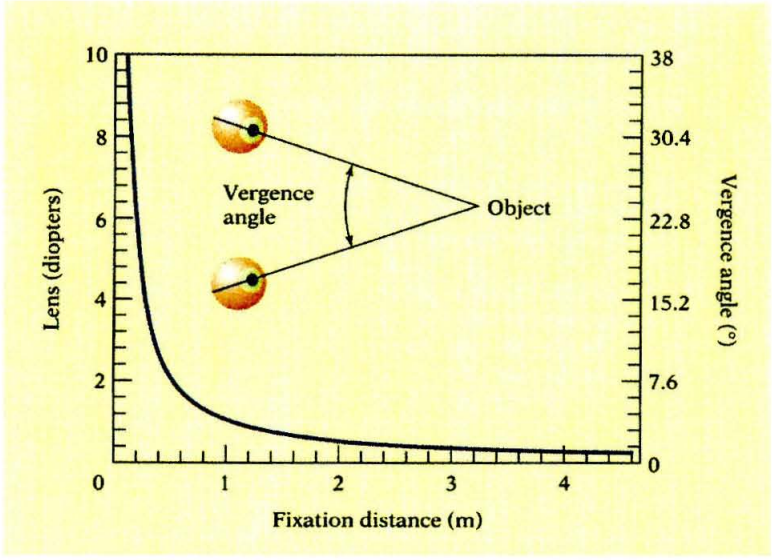


Figure 1.4: Accommodative Power and Convergence Angle as a Function of Distance Changes in the accommodative state of the ocular lens and the vergence angle between the eyes as a function of the fixation distance between the viewer and the object. The curves for these two functions are identical. Note that for distances greater than 1 meter there is little change in accommodation power or vergence angle. This means that these cues will be of little use in determining the distance of objects greater than 1 meter away, and that, as proposed by Descartes, the visual system must rely on cues that are strongly dependent on learning and experience. Taken from [7, Evolving Brains, Allman].

Lesion studies have indicated that visual cortical area V4 is important in integrating distance cues with retinal size information to arrive at accurate judgments of object size. Electrophysiological studies in visual cortical area V4 have shown that distance cues, both visual and motor, modulate the size tuning curves of neurons. This is illustrated in figure 1.5.

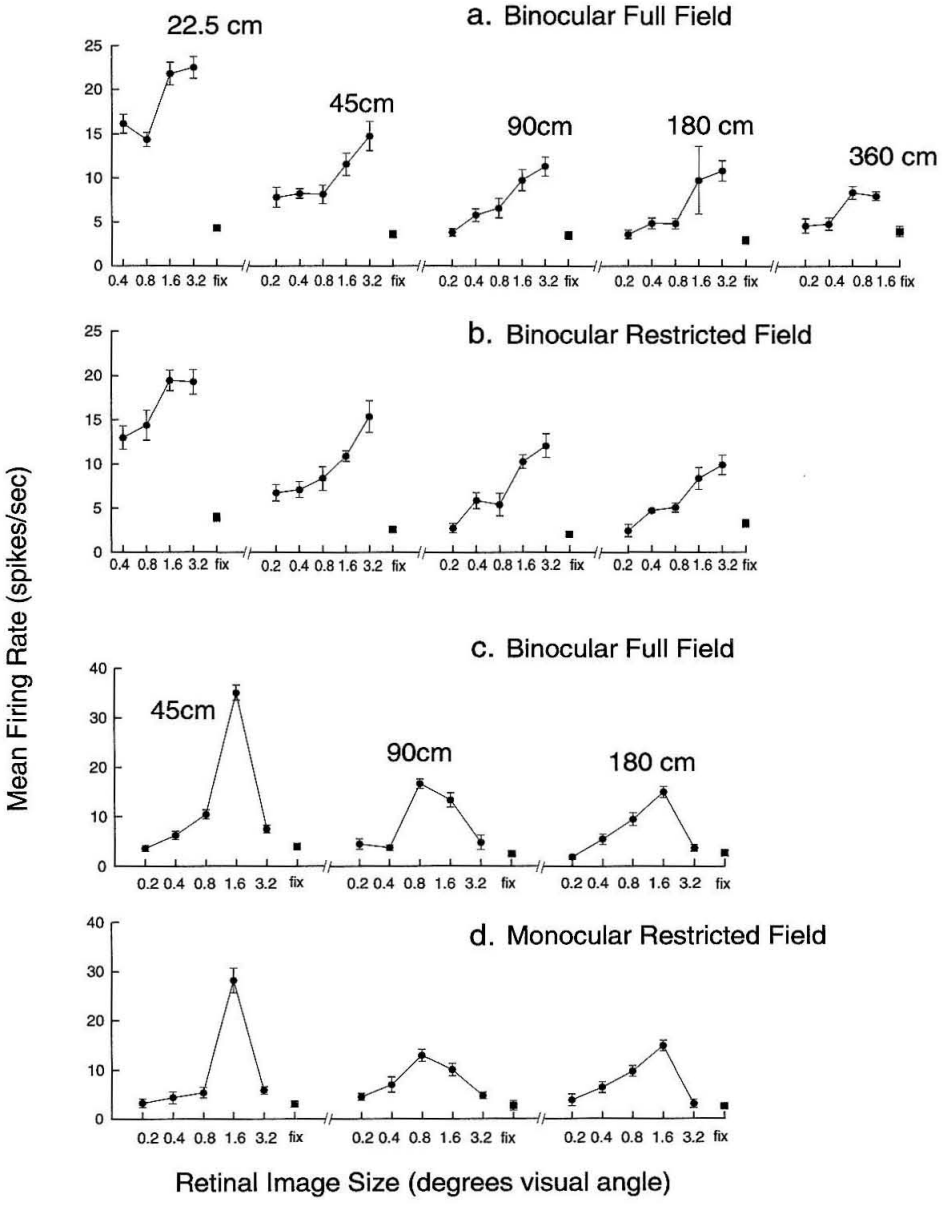


Figure 1.5: Distance Modulated Cells

Size tuning curves of V4 cells showing modulation with respect to distance which persists even when pictorial cues are removed. Figure from *Representation of Three-Dimensional Space in Primate Visual Cortex*, Richard M. Jeo, Caltech Thesis 1998[43].

1.2.2 Sources of Eye Position Signals in Visual Cortex

There are two types of extra-retinal eye position signals, categorized according to their point of origin: those that originate from receptors in the eye muscles; and those from areas commanding the eye position. We will refer to the two as afferent feedback and efference copy respectively. Since the pathways carrying each of these types of information consist of overlapping recurrent networks of cortical areas and sub cortical nuclei, the question of where the eye position signals originate and which of these two types of information is being conveyed may be ill-posed.

While the more restricted anatomical question of which areas provide a direct input which could carry this type of information can be addressed (and will be discussed in detail in the next section), the data from psychophysical approaches is much less conclusive and often produces conflicting results. Psychophysical experiments provide indirect evidence that both afferent and efferent inputs contribute to the perception of eye position. Studies in which eye muscles are temporarily paralysed, effectivly removing afferent feedback from eye muscles while leaving efference copy signals intact, support a role for efference copy in determining perceived position[14]. But complementary studies in which the eyes are not moved, eliminating efference copy, but eye muscle vibration produces false afferent feedback, also provide support for the role for afferent feedback into the visual system[82]. There is evidence that proprioceptive feedback indicating head and body orientation are also influential on observer's perception of visual direction[68].

Efference Copy from the Oculomotor Command System

The frontal eye fields (FEF) are an important component of the cerebro-ponto-cerebellar pathway involved in governing voluntary eye movements, including vergence and ocular accommodation[28]. In addition to its connections with the pontine nuclei and its well known projections within frontal cortex, the FEF has projections to many posterior visual cortical areas, including those studied in this thesis. The FEF is broadly divided into an area governing small saccades (sFEF)[4] and an area governing large saccades (lFEF), with sFEF providing a much heavier projection to posterior visual areas than lFEF. Both areas are topographically organized according to saccadic amplitude, and loosely maintain this

[4]Small saccades are those less than 10^o in amplitude, which are by far the most common, directing gaze to conspicuous and informative features of a scene.

organization in their projections to visuotopic cortical areas. In general, projections from lFEF terminate in areas with large and eccentric receptive fields[5], whereas sFEF projections terminate in areas with small centrally located receptive fields. Of particular interest in our study are the projections from sFEF to areas V2-V4 of the lunate sulcus and V3-V4 along the medial wall of the infereior occipital sulcus. There is a bilaminar pattern of termination of FEF projections to most of the posterior cortical areas, much different from the columnar pattern, in which termination is more evenly distributed throughout all cortical layers, found in frontal lobe projections. Usually termination is restricted to layers 1 and 5/6 with greater density in layer 1 and a looser meshwork in layers 5/6. The columnar pattern was found most notably in area 7a/LIP, which is also the recipient of the heaviest of the posterior FEF projections. These two different laminar termination patterns may have their origin in two different populations of projection neurons within the FEF; the bilaminar pattern originating in FEF cells primarily in layer 5/6; the columnar pattern originating in large FEF cells of layer 3[77].

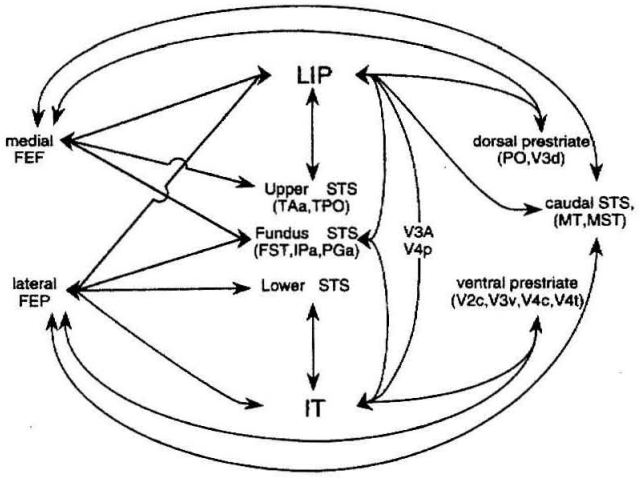


Figure 1.6: Connections of Frontal Eye Fields

Diagram showing anatomical connections between Parietal, Inferotemporal, and Frontal cortices, Diagram from [17, Bullier et al.]

While many cells within FEF exhibit transient responses during eye movements, there

[5]Such as at the mouth of the calcarine sulcus at the border of V1 and V2

is a population of cells that display a tonic firing rate related to vergence angle and accommodation. Cells have been identified with activity specifically linked to the near-response or to the far-response[28]. Since we are measuring firing rates during a period 500-1000msec after the production of a saccade, the cells which exhibit a tonic response may play a role in the modulation found in the studies discussed in this thesis.

While the connections from FEF to posterior visual areas provides a means by which commanded eye position signals could influence low level visual processing, there are also reciprocal connections from posterior visual areas to FEF which may be important in modulating the formation of eye position commands. In particular, lateral FEF receives a major part of its input from the ventral part of prestriate and inferotemporal cortex (V3v,V4,TEO) which probably signals feature attributes to be used for selecting the target for eye movements [17]. The connections between frontal, parietal, and inferotemporal cortex appear to be organized as a network of interrelated areas emphasizing central vision, small saccades, and form recognition. Psychophysical studies dating back to the work of Yarbus demonstrate that patterns of fixations depend both on features in the visual scene ("bottom up processing") and on the questions one is trying to answer from the information contained in the scene ("top down processing")[86; 49]. The common finding that eye movements are directed to features of faces such as eyes, nose, and mouth is likely to rely upon information from ventral regions involved specifically with face processing relayed to FEF. Eye-movement theories of optical illusions have established a relationship between distortion of perception and distortion of eye movements during perception[66]. Enright has found that accommodative vergence varies with the implied depth of the point fixated when viewing a painting with strong perspective cues[24]. Electrophysiological stimulation studies lend additional support to the notion that activity in ventral visual areas influences eye-movements. Stimulation of the transition area between occipital and temporal cortices elicits the three components of the near response, accommodation, pupillary constriction, and convergence [42]. The stimulation sites from which these responses were elicited correspond with areas found to have anatomical connections with FEF[81]. PET studies of the near response provide further evidence of cortical processes producing increases of activity in posterior structures (occipital, cerebellar, and temporal) and with activity decreases in frontal and parietal regions preceding the ciliary motor command. These studies suggest a dynamic and reciprocal functional connection between the accommodation and the visual

search/visual attention systems that share premotor circuitry[65].

Physiology of Dorsal Visual Cortical Areas

Physiological experiments in dorsal cortical areas have examined in great detail the neural representation of visuospatial relationships, including distance and angle of gaze. Extensive work in parietal areas has shown that gaze angle and object depth modulates the gain of responses in parietal areas. The primary mechanism for gaze modulation is reported to be linear gain modulation of neural responses (for reviews see [10]).

Sakata et al. [1980] recorded from area 7a of posterior parietal cortex in macaque monkeys that were trained to fixate on a movable screen that varied in distance from the monkey. They found that the neurons fired when the monkey fixated on the target and that for visually responsive neurons in area 7a, the magnitude of the response was related to fixation distance. Sixty-five percent of the depth sensitive neurons preferred closer fixation distances and 29% preferred farther distances.

Representation of Distance in Primary Visual Cortex

Distance information is represented as early as primary visual cortex and also in parietal cortex. Recordings from V1 in awake monkeys trained to fixate on a spot on a monitor that could be moved to different distances have measured responses to binocular disparity at different viewing distances. The stimuli were scaled so that retinal dimensions were kept constant. The primary result was that the magnitude of these responses was modulated by viewing distance for about 80% of the neurons studied. Angle of gaze is also represented as early as primary visual cortex[43; 84]. Although this study was performed in cats, it is likely that direction of gaze is represented in V1 of primates.

1.2.3 Amygdalar Inputs to Visual Cortex

The nuclei of the amygdala play an important role in the control and modulation of behaviors associated with emotional and visceral reactions, particularly organization of appropriate response to threat or attack. Many of the complex series of changes accompanying bilateral damage to the amygdalar complex, referred to as Kluver-Bucy Syndrome, can most easily be explained on the assumption that the ventral visual processing stream has

become disconnected from the system which attaches an appropriate motivational tag to percepts. For example, monkeys with amygdalar lesions are insensitive to visual stimuli that normally arouse intense fear. Another profound effect of damage to amygdala is to make it difficult for reinforcing stimuli, whether positive or negative, to become established or recognized [83]. There is growing evidence that the role of amygdala in computing an affective motivational tag given a sensory input, and relaying this tag back to the sensory areas providing the input, contributes to the association of reward with previously neutral stimuli. In humans, the amygdala appears to play a general role in guiding preferences to visual stimuli that are normally judged to be aversive or to predict aversive consequences. This function may be especially critical in regard to the judgment of social stimuli such as faces, as evidenced by the specific deficits in the recognition of affective facial features relating fear, and spared recognition of identity[2; 3]. More broadly, Amaral suggests that amygdalo-cortical projections might have a role in modulating cortical processing based on the motivational or emotional state of the organism[9]. LeDoux has noted that the projections from amygdala to cortex are considerably heavier than from cortex to amygdala. Amygdala projects to primary and secondary visual processing areas from which it does not receive inputs. Visual inputs to amygdala come primarily from much higher level visual areas in temporal cortex such as TE and TEO. With this architecture, activation of amygdala by complex visual stimuli could result in feedback to early visual processing areas, altering the processing of subsequent stimuli[50].

LeDoux has also pointed out that a potential purpose for a visual input from thalamus to amygdala, in addition to the much more highly processed input from temporal cortex, is simply to provide a fast acting early warning of potential danger. We would like to add to this the idea that an early warning signal activating amygdala might trigger a potentiating input from amygdala to early visual cortical areas which would be useful in preparing processing in these areas for responding to danger as illustrated in figure 1.7.

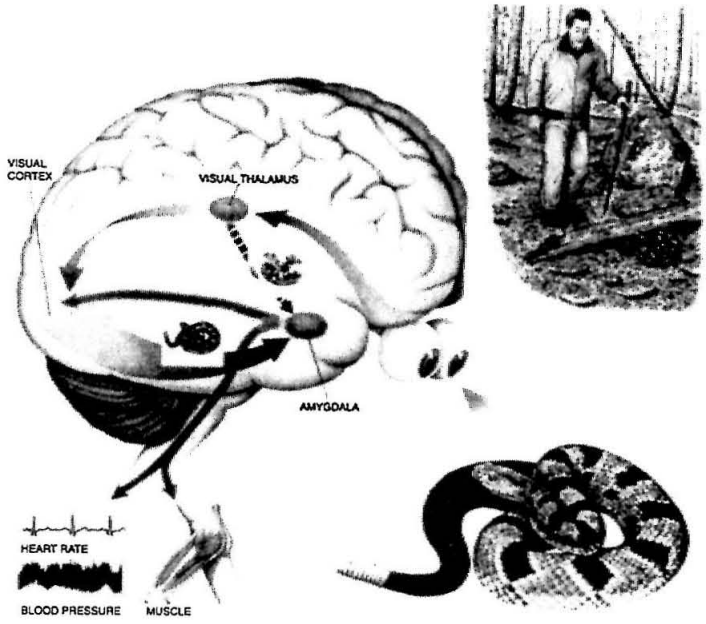


Figure 1.7: Anatomy of Fear Response

Diagram showing a scenario in which an unexpected danger is encountered, and how amygdala organizes a response, including sending a signal to ventral visual cortical areas. Diagram modified from LeDoux.

Evidence of the capacity for signals from amygdala to potentiate or stimulate visual system processing comes from electrical stimulation of amygdala[6], which frequently illicit complex visual hallucinations. The hallucinations have been interpreted as the result of neocortex attaching a significance signal to random cortical activity[31].

1.2.4 Neuromodulatory Inputs to Visual Cortex

Neuromodulatory inputs to cortex consist of diffusely projecting, widespread afferents which use one of several monoamines as neurotransmitter[7]. The broad spatial domain of these projections and the relatively long time course of monoamine actions makes these systems ideally suited for a role in regulation of activities that involve large areas of neocortex, such

[6] In patients with temporal lobe epilepsy

[7] Acetylcholine will be included.

as vigilance, attention, affective state changes, and mood. Each monoamine projection originates from a separate nuclear complex at different levels of the neuraxis. The tangential termination patterns of these systems, in which single axons may innervate different functional cortical areas, differ fundamentally from the termination patterns of thalamocortical and corticocortical afferents.

- NA innervation arises from the locus coeruleus. Tecto-pulvinar-extrastriate structures are more densely innervated than geniculostriate and inferotemporal structures. The preferred target of NA innervation is pyramidal cell dentrites of layers III, V, and VI.

- 5-HT innervation arises from the dorsal and median raphe nuclei. The seroternergic neurons innervating primary visual cortex are separate from those that innervate prefrontal, motor and somatosensory areas. There are two classes of 5-HT fibers, very fine and larger caliber, each having its own regional and laminar preferences. These fibers show a strong preference for layer IV in area V1, where the innervation is among the densest of all neocortical areas. Relatively small diameter distal dendrites of both pyramidal and non-pyramidal neurons are the primary target of serotonergic input.

- DA innervation arises from the substantia nigra/ventral tegmental area cell groups. There is only a very sparse projection to area V1, where it is limited to layer I; V2 receives slightly more innervation primarily in layers I and V/VI; a projection of intermediate density to temporal visual areas terminates in all layers except IV.

- ACh innervation arises from the nucleus basalis of Meynert. There is a substantial innervation of V1 and other primary sensory and motor areas, and a less dense innervation of visual association areas. In V1, layer I receives the most dense innervation followed by layers II and III.

It is notable that FEF inputs to the visual areas we are recording from have the same laminar pattern of termination (primarily in layers I and V/VI) as the dopaminergic input. Amygdalar inputs also terminate primarily in layer I. The dopamenergic and amygdalar inputs to these areas are of particular interest in this context since they have been strongly implicated in positive and negative reinforcement learning respectively.

Tonic activity of cortico-cortical inputs could have effects on post synaptic sites which last as long as the duration of the tonic activity. Hence this type of input could have

potentially long duration effects, as neuromodulatory inputs do, but have the advantage of being able to turn on and off rapidly, and have spatially more specific connections.[8] Although the inputs from FEF, amygdala, and DA neurons act primarily on distal dendrites in layer I, mechanisms such as distance-dependent scaling of synaptic strength, which are found in cortical pyramidal cells, can make the effect of these distal synapses as strong as those closer to the soma [52].

1.2.5 Other Potential Sources of Inputs to Visual Cortex

Central thalamic nuclei of primate contain neurons related to vergence and ocular accommodation that primarily carry signals related to the motor commands for vergence and accommodation. These nuclei have projections into visual cortical areas and hence might be a potential source of the extra-retinal eye position signals we observed[87].

Based on clinical neuropsychological studies, researchers have suggested that the hippocampus and medial temporal lobe structures are important for encoding what has been variously described as contextual, configural, spatio-configural, or relational information. It has been demonstrated that implicit memory [9] for contextual visuospatial information facilitates perceptual processes such as visual search. The lack of such contextual cueing effects in amnesic patients' in the presence of intact perceptual/skill learning, suggests that medial temporal lobe may function to bind contextual cues in the environment[21] and provide this contextual information to visual cortical areas for use in forming associations. In serving such a function, these areas may be important to some forms of priming (sensory conditioning).

[8]The fact that equivalent changes in electrical membrane properties can be produced either by changes in the biophysical properties of membrane channels, or by appropriate direct electrical stimulation of the membrane, has been exploited in an experimental technique called dynamic clamping[73].

[9]Implicit memory is most succinctly described as memory not accessible to consciousness

1.3 Experiments

1.3.1 Methods

General Methods

All surgical, training and neurophysiological recording procedures conformed to the National Institutes of Health and USDA guidelines for animal research, and were carried out under a protocol approved by the Caltech Institutional Animal Care and Use Committee (IACUC).

Training and Surgery

Two adult male macaque monkeys (one *Macaca mulatta* and one *Macaca fascicularis*) were used in this study. The monkeys were trained to sit in a standard monkey chair and fixate a small spot on the computer monitor for a juice or water reward delivered by a device capable of dispensing up to four different types of liquid with up to 0.01 ml accuracy (Mike Walsh, Caltech Biology Electronics Shop). Prior to training, a stainless steel head post was implanted to permit head restraint for fixation training and recording. The head post was fixed to the skull using orthopedic straps and bone screws (Synthes, USA) under sterile conditions and general anesthesia (xylazine 0.5 mg/kg, ketamine 10 mg/kg). Post operative analgesic drugs (buprenorphine, codine, acetaminophen) were administered for several days following the surgery. Fixation was monitored with a non-invasive infrared video-based eye tracker (ISCAN, RK-716PCI). Following fixation training, a second aseptic surgery was performed to implant a recording chamber (Caltech Central Engineering) over a craniotomy to allow controlled and sterile insertion of microelectrodes. The placement of the chamber was determined by cranial and vascular landmarks, and was designed to give access to parafoveal areas of V1, V2, and V4. The correct placement of the chamber and location of the lunate sulcus was verified through exploratory receptive field mapping. The location of the chamber limited access to those neurons whose receptive fields were predominantly located parafoveally in the lower left quadrant of the visual field.

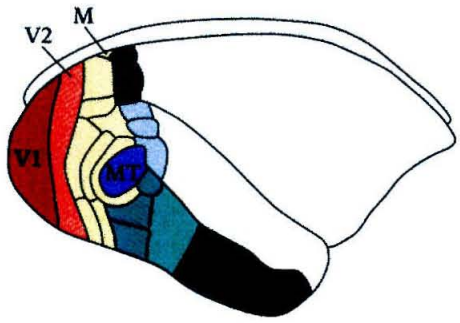


Figure 1.8: Illustration of Brain

Illustration of primate brain with different visual areas color coded.

The chamber and implant margins were monitored for infection, and were cleaned daily with saline and dilute chlorhexidine diacetate (0.05%). The chamber was filled with inert sterile oil (either heavy silicon or mineral oil) during and between recording sessions. Sterile opthalmic antibiotic ointment (Bactracin-Neomycin-Polymixin, Gentacin or Chloramphenicol) was used as necessary (for 10-14 day periods), to inhibit bacterial growth in the recording chamber.

Recording

To record the activity of single neurons, the intact dura was penetrated with sterile glass insulated platinum-iridium microelectrodes (1-4 $m\Omega$), using a stepping motor micro drive (Herb Adams, Caltech Central Engineering). The location of the penetration was set in polar coordinates using the chamber opening as the frame of reference (radius 0mm being the center of the opening, and angle 0 degrees being determined by a notch on the right lateral side of the opening). The electrode signal was amplified by a preamplifier at the head stage (Mike Walsh, Caltech Biology Electronics Shop) and then band pass filtered before being digitized. The analog to digital conversion was performed by a PCI-1200 National Instruments data acquisition board in a dual processor Intel PIII Windows NT based machine. Acquisition software was written with Labview5.1 (National Instruments). Single neurons were isolated using a window discriminator with 5 parameters written in Labview (Yuxi Fu). Spike times were recorded to 1ms resolution.

Visual Stimuli

Setup and Receptive Field Mapping

A computer monitor (SGI Graphic Display Monitor) was mounted on a precision computer controlled positioning device (Industrial Devices Corp.) 87cm long. Electro-optic sensors marked three positions on the positioning device track which were points at which the center of the surface of the monitor was at 22.5cm, 45cm, and 80cm from the monkey's eyes.

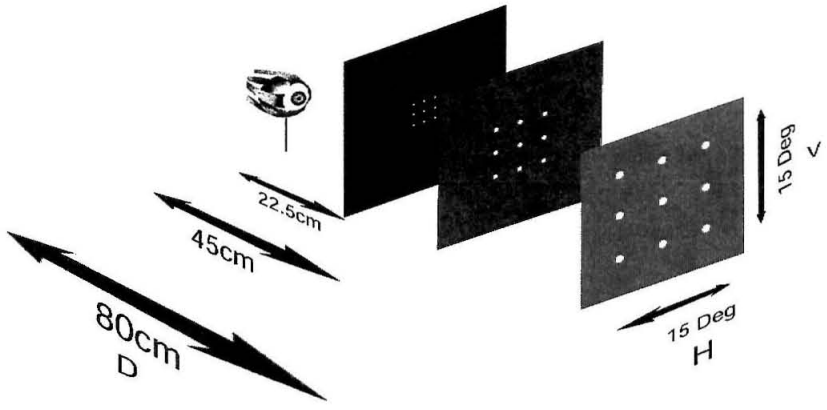


Figure 1.9: Experimental Paradigm

Subjects were required to fixate on a spot that might appear at any one of 27 different positions. The possible positions of the fixation spot consisted of three horizontal positions, three vertical positions and three distances.

The monitor was used in its highest resolution mode (1280x1024 pixels) with a refresh rate of 75Hz. Stimuli were generated on an SGI O2 using graphics programs written in Python and utilizing native SGI OpenGL. Whenever applicable, antialiasing routines were used to reduce pixellation effects. For each subject, an extensive calibration was performed to determine the center of the monitor (the intersection between the line passing through the point midway between the eyes and perpendicular to the monitor surface). This calibration was repeated regularly during the time period over which these experiments were conducted. The monkey's chair and the positioning device were aligned so that the motion of the monitor was perpendicular to the plane passing through the monkey's eyes and perpendicular to the ground.

The subjects viewed the monitor through an aperture constructed from a matte black material placed approximately 8cm from the eyes. The aperture masked off all of the visual environment except the center of the monitor screen, even at the farthest distance. This was verified before the start of an experiment in two ways: using rigid rods to determine the line of sight, and by testing the limits of the subjects field of view behaviorally by having the monkeys fixate targets at different locations on the screen.

Stimulus onset and offset events were synchronized to the data collection using a trigger spot appearing briefly at the leftmost edge of the monitor during the onset and offset image frames. This spot was detected by a small photodiode affixed to the leftmost edge of the monitor, and this trigger signal was sent as an analog signal to the data acquisition board. The rising edge of this signal was used as the stimulus event onset time. The trigger spot was not visible to the subject as it was masked by the aperture.

When a single neuron was isolated, the optimal receptive field characteristics were estimated by hand at one or two viewing positions (usually the central position at 80cm and/or 45cm) using bar stimuli. The location, size, aspect ratio, orientation, brightness, speed, and direction of motion of the bar were adjusted so as to produce the maximal response from the isolated neuron. During the experiment, the bar stimulus to the receptive field was presented statically (flash on- flash off) so that the temporal characteristics of the neural responses could be more easily assessed.

Experimental Paradigm

A single successful experimental trial consisted of the following events: fixation spot goes on; subject acquires fixation within 50 msec and maintains fixation for 500 msec; bar stimulus comes on in the receptive field and stays on for 1500-2000 msec; bar stimulus goes off; fixation spot goes off. If, after acquiring fixation, the subject maintains fixation until the fixation spot goes off, then he is given a juice or water reward. If, at any point after acquiring fixation and before the fixation spot goes off, the subject breaks fixation, then the trial is immediately aborted, the screen is blanked, and there is a short pause interval before the beginning of the next trial.

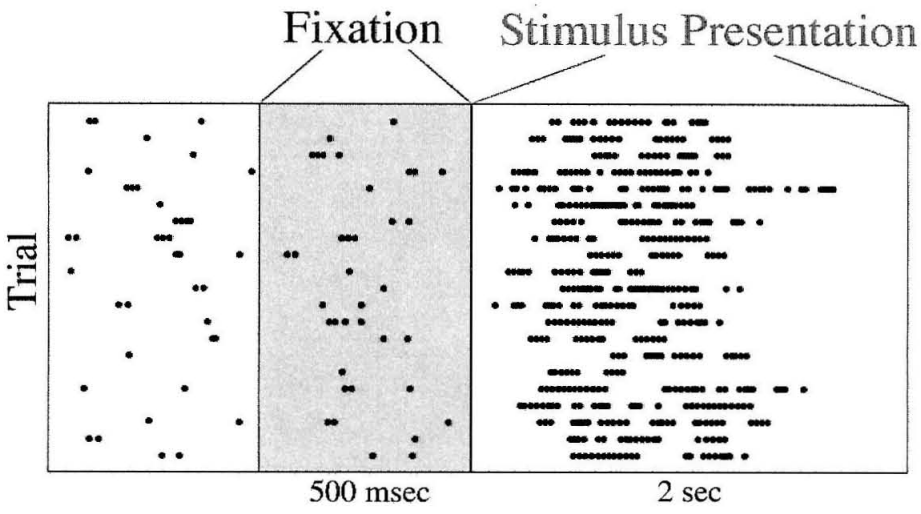


Figure 1.10: Trial Structure

Responses during fixation only and during stimulus presentation over the receptive field of the neuron were measured each trial, and mean spike rate for both portions of the trial were calculated. The monkey was required to maintain fixation throughout the duration of the entire trial.

Before each trial an experimental condition is selected at random from the set of 27 possible (h, v, d) triplets. Each experimental condition is repeated 10 times (in random order), so that there are a total of 270 trials per experiment. The positioning device is then commanded to move the monitor to the position determined by the value of d selected for this trial. Sizes of stimulus elements and distances between stimulus elements are scaled with distance using the subject's center of monitor is the origin. This is done so that the visual stimulus falling on the retina is constant throughout the experiment, regardless of experimental condition.

1.3.2 Data Analysis

Area Classification

Cells were assigned to a visual cortical area based on receptive field position, size, and properties, and position relative to the lunate sulcus. In the absence of histological classification,

we combine $V1$ and $V2$ for quantitative analysis.

Tests for modulation

It is well known in the literature on cortical physiology that there is a linear relationship between the mean firing rates and variance in the firing rates among cortical cells, and we found that this observation holds true for our data set, both taken as whole or as individual cells. This type of correlation between mean firing rate and variance in the firing rates violates the equality of variances (or homoscedasticity) assumption of the standard anova model [76] which is typically used to determine if the difference in mean firing rates under different experimental treatments are statistically significant.

There are at least two ways to perform statistical significance tests in this situation. The standard treatment of data in which means and variances are positively correlated is to logarithmically transform the data before performing an ANOVA test. Another possibility, is not to assume anything about the distributions and perform a non-parametric test such as the Kruskall-Wallis Rank Sum test. Both these analysis techniques were applied and the two analyses were in such close agreement that we will only quote the results of the more standard log transform anova analysis. Three-way ANOVAs were performed on both the log transformed data from the stimulation period and the log transformed data from the fixation period to determine if there were significant modulations of the mean firing rate with respect to each of the experimental variables H,V, and D or any combination of them. In all tests of significance a p-value of 0.01 was used as criterion threshold.

Fractional Gain

The magnitude of modulation of the mean response with respect to each of three dimensions was quantified by calculating the fractional gain between the highest and the lowest mean response values. We will denote the mean firing rate for the 10 repetitions of experimental condition (h, v, d) by $M_{(h,v,d)}$, which is calculated by dividing the spike count by the stimulus duration. The mean spike rate was also calculated for the fixation only period. Using this notation, we define the maximum and minimum mean firing rates with respect to each of the three experimental variables as follows:

$$M_{max}(v,d) \equiv \max_{h}(M_{(h,v,d)}) \qquad M_{min}(v,d) \equiv \min_{h}(M_{(h,v,d)}) \tag{1.1}$$

$$M_{max}(h,d) \equiv \max_{v}(M_{(h,v,d)}) \qquad M_{min}(h,d) \equiv \min_{v}(M_{(h,v,d)}) \tag{1.2}$$

$$M_{max}(h,v) \equiv \max_{d}(M_{(h,v,d)}) \qquad M_{min}(h,v) \equiv \min_{d}(M_{(h,v,d)}) \tag{1.3}$$

The definition of the fractional gain values can now be formulated as follows:

$$FG_h(v,d) \equiv \frac{M_{max}(v,d) - M_{min}(v,d)}{M_{max}(v,d)} \tag{1.4}$$

$$FG_v(h,d) \equiv \frac{M_{max}(h,d) - M_{min}(h,d)}{M_{max}(h,d)} \tag{1.5}$$

$$FG_d(h,v) \equiv \frac{M_{max}(h,v) - M_{min}(h,v)}{M_{max}(h,v)} \tag{1.6}$$

Each of these functions for fractional gain gives 9 fractional gain values since there are three possible values for each of their two arguments. These 9 values can be summarized as a single value, the mean:

$$FG_h \equiv \mathop{\mathbf{E}}_{(v,d)}[FG_h(v,d)] \tag{1.7}$$

$$FG_v \equiv \mathop{\mathbf{E}}_{(h,d)}[FG_v(h,d)] \tag{1.8}$$

$$FG_d \equiv \mathop{\mathbf{E}}_{(h,v)}[FG_d(h,v)] \tag{1.9}$$

The lowest possible fractional gain value is 0.0 which indicates that the mean response rate was unaffected by a change in the dimension in question. The highest possible value of 1.0 indicates that responses were absent for at least one value of the dimension in question.

Modulation Index

For each of the dimensions H, V, and D cells were classified into three categories.

- H
  - Leftness: monotonically decreasing with respect to h

  - Non-monotonic: non-monotonic with respect to h
  - Rightness: monotonically increasing with respect to h

- V
  - Downness: monotonically decreasing with respect to v
  - Non-monotonic: non-monotonic with respect to v
  - Upness: monotonically increasing with respect to v

- D
  - Nearness: monotonically decreasing with respect to d
  - Non-monotonic: non-monotonic with respect to d
  - Farness: monotonically increasing with respect to d

A modulation index for each the dimensions H,V, and D was calculated as follows:

$$Mod_h \equiv FG_h \times Class_h \quad \text{where} \quad Class_h \equiv \begin{cases} 1 & \text{if cell is a Rightness cell,} \\ 0 & \text{if cell is a Non-monotonic cell,} \\ -1 & \text{if cell is a Leftness cell.} \end{cases}$$

$$Mod_v \equiv FG_v \times Class_v \quad \text{where} \quad Class_v \equiv \begin{cases} 1 & \text{if cell is a Upness cell,} \\ 0 & \text{if cell is a Non-monotonic cell,} \\ -1 & \text{if cell is a Downness cell.} \end{cases}$$

$$Mod_d \equiv FG_d \times Class_d \quad \text{where} \quad Class_d \equiv \begin{cases} 1 & \text{if cell is a Farness cell,} \\ 0 & \text{if cell is a Non-monotonic cell,} \\ -1 & \text{if cell is a Nearness cell.} \end{cases}$$

1.4 Results

These experiments were designed to examine the effect of point of regard on neural responses as measured by mean firing rate of the neuron during both the fixation only period (symbolized FO) and the stimulus presentation period (symbolized S). A total of 88 cells (41 in $V1$ and 47 in $V4$) were recorded from in two monkeys, while they performed a fixation task. These visual areas belong to early and intermediate stages of visual cortical processing along the ventral pathway. We found that 85% of the cells recorded from had a statistically significant amount of modulation with respect to at least one of the experimental variables H,V, or D. Figure 1.11 and figure 1.14 show examples of the type of modulation encountered.

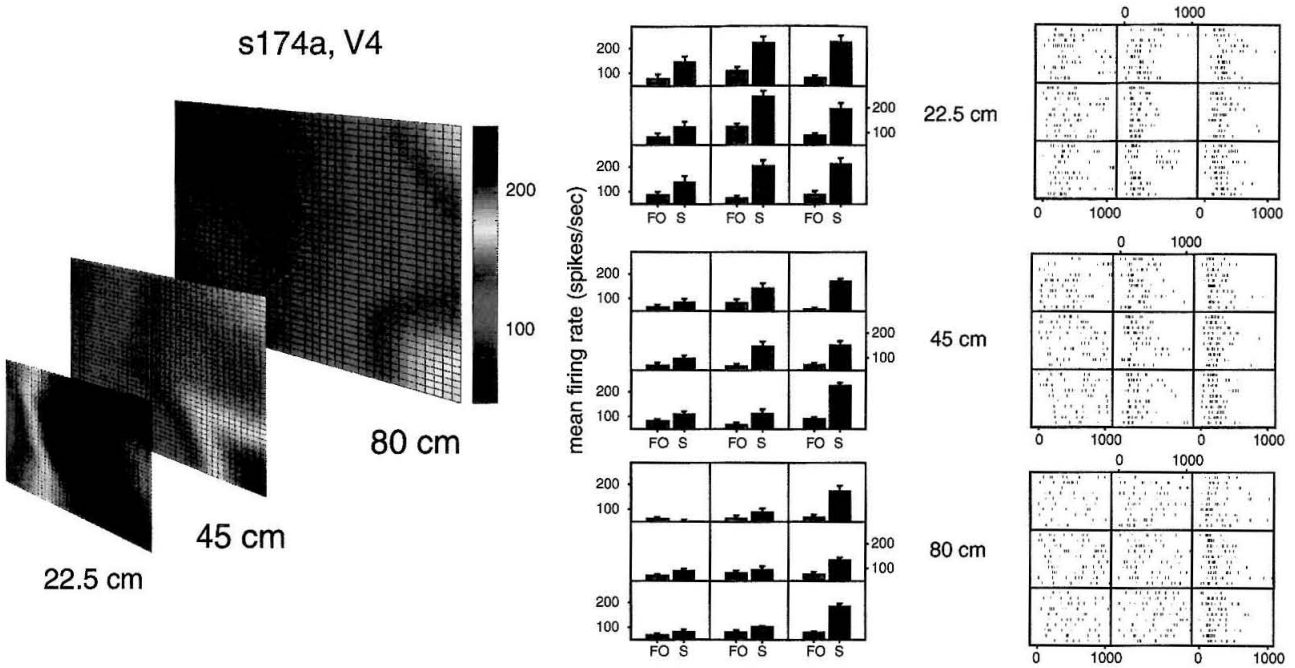


Figure 1.11: Example Cell: Three Representations

Three different representations of eye position modulation data from a single cell showing modulation with respect to both H and D. In the center, a bar graph representation where the height of the bars represents the mean firing rate during stimulation (S) and fixation only (FO) periods. The bar graphs are organized into three panels, one for each viewing distance, each panel consisting of a three by three matrix corresponding to the three by three possible (h, v) pairs. On the right, a spike raster representation of the data for the stimulation period. Each horizontal line represents a single trial of spike data. The horizontal position of the spikes represents the time since stimulus onset in milli-seconds. On the left, an interpolated color map representation of the same data, where color represents mean firing rate during the stimulation period only. The three image planes represent the three viewing positions with the smallest image corresponding to the closest distance and the largest plane corresponding to the farthest distance. Position within the image planes represents the (h, v) position of the fixation spot.

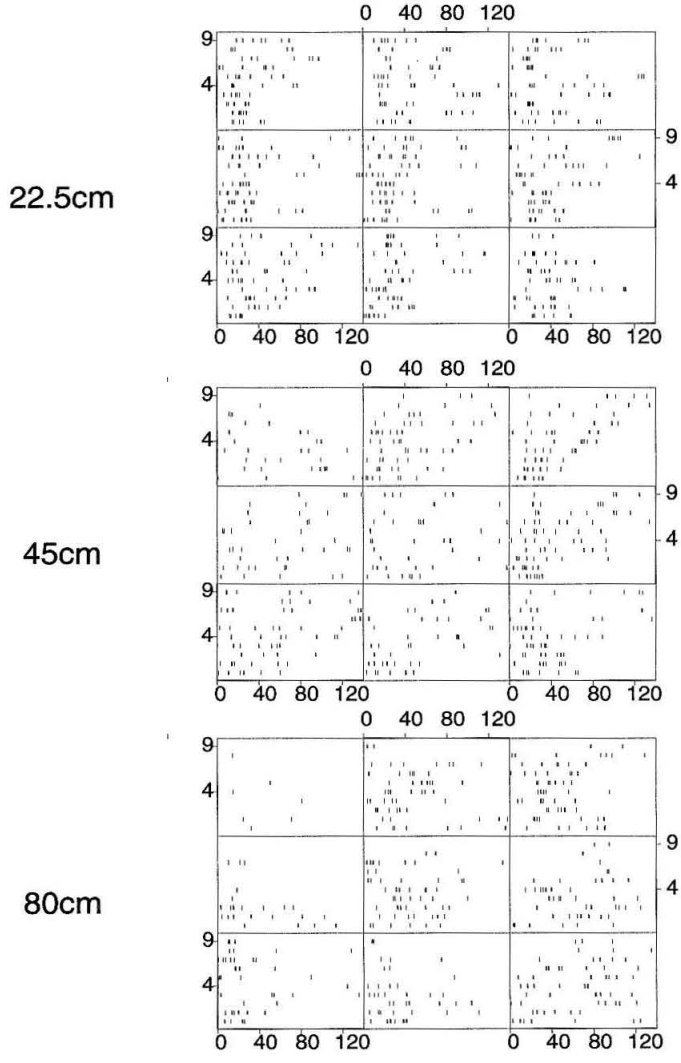


Figure 1.12: Raster Representation of Example Cell
This V4 cell shows a nearness preference and at the far distance a horizontal preference for direction to the right.

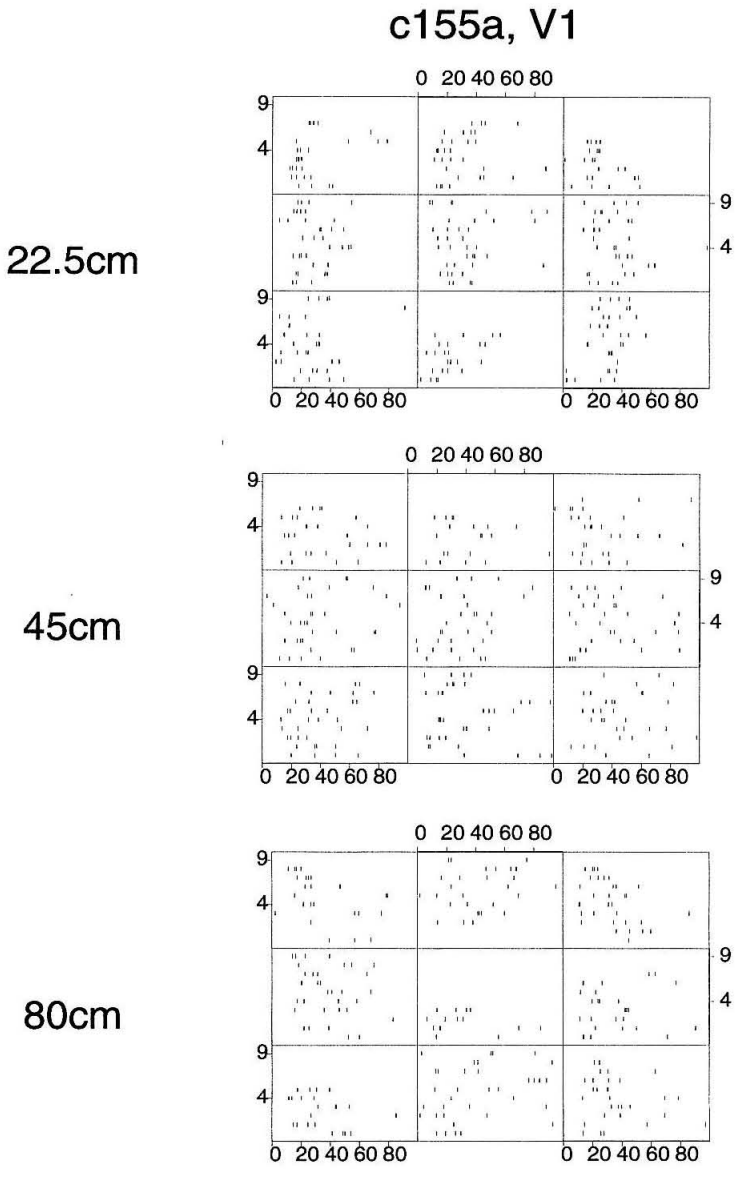


Figure 1.13: Raster Representation of Example Cell
This V1 cell shows a complex non-monotone modulation with a clear nearness component.

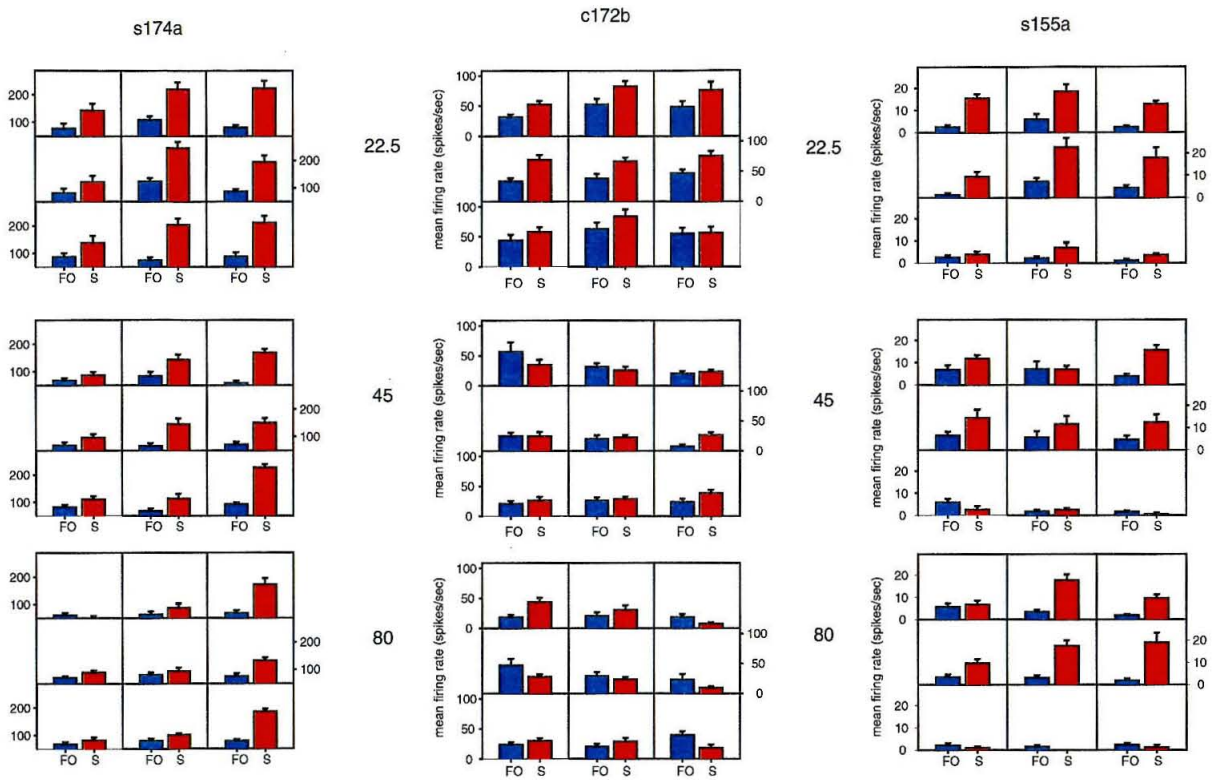


Figure 1.14: Three Examples

Examples of different types of modulation of mean firing rate with respect to point of regard. The three columns represent three different cells. The three rows represent the three possible viewing distances. Each panel is divided into nine bar graphs representing the nine possible (h, v) pairs. FO is the mean firing rate during the Fixation Only period of the trial. S is the mean firing rate during the Stimulation period of the trial. The leftmost figure represents a cell which exhibits mostly modulation with respect to H but also a little modulation with respect to D. The middle figure is an example of modulation with respect to D. This cell is a nearness cell. The rightmost figure is an example of modulation with respect to V. This cell is an upness cell.

1.4.1 Demographics of Modulation Effects

Each cell in our sample population was tested for modulation with respect to H,V,D and all possible interactions between these variables using a MANOVA analysis. The percentage of the population having each of these different types of modulation is shown in figure 1.15 below. The distribution of the different types of modulation varies significantly both with respect to visual cortical area (V1/V4) and with respect to trial period (FO/S).

In all cases the largest population of cells were those modulated with respect to distance.[10] In all cases there were few cells which were modulated with respect to both H and V.[11] The distribution of the different types of modulation has a strikingly similar shape for the fixation only period and the stimulation period within each area. The similarity in the shapes of the distributions during these two trial periods is evidence that similar modulatory mechanisms are operating during the fixation only period and during the stimulation period, and that part of the modulation found in the stimulation period may be accounted for by modulation already present during the fixation only period[12] The amount of modulation during the stimulation period is larger than that found during the fixation period in both V1 and V4.

There are some notable differences between the distribution for area V1 and the distribution for area V4. In V1 there is a paucity of modulation with respect to V, a result which confirms earlier findings[79], while in $V4$ the amount of modulation with respect to V is comparable to the amount of modulation with respect to H[13].

The modulation of each cell can be classified according to whether the mean firing rate is monotonically increasing, monotonically decreasing, or neither, with respect to the experimental variables. The measures $(Class_h, Class_v, Class_d)$ used for this purpose are described formally in the section on data analysis. The distribution of these different classes

[10] An analysis of the values of the fractional gains revealed that, although the average magnitude of modulation with respect to D was marginally larger than those with respect to the other dimensions (statistically significant), the slightness of this difference leads us to believe that it is the more widespread distribution of this type of modulation, rather than its greater strength, which accounts for its greater representation in the population.

[11] This may be indicative of an independence of the sources of the signals producing the H and V modulations. Likewise the coupling between the H and D modulations found in V1 may be indicative of a common source for the modulation with respect to these two parameters.

[12] This will be discussed in greater detail below.

[13] This may be an indication that information regarding V is not available to V1 and only enters the visual processing stream at the later stage of V4. This is further evidence of the independence of the sources of information regarding H and V.

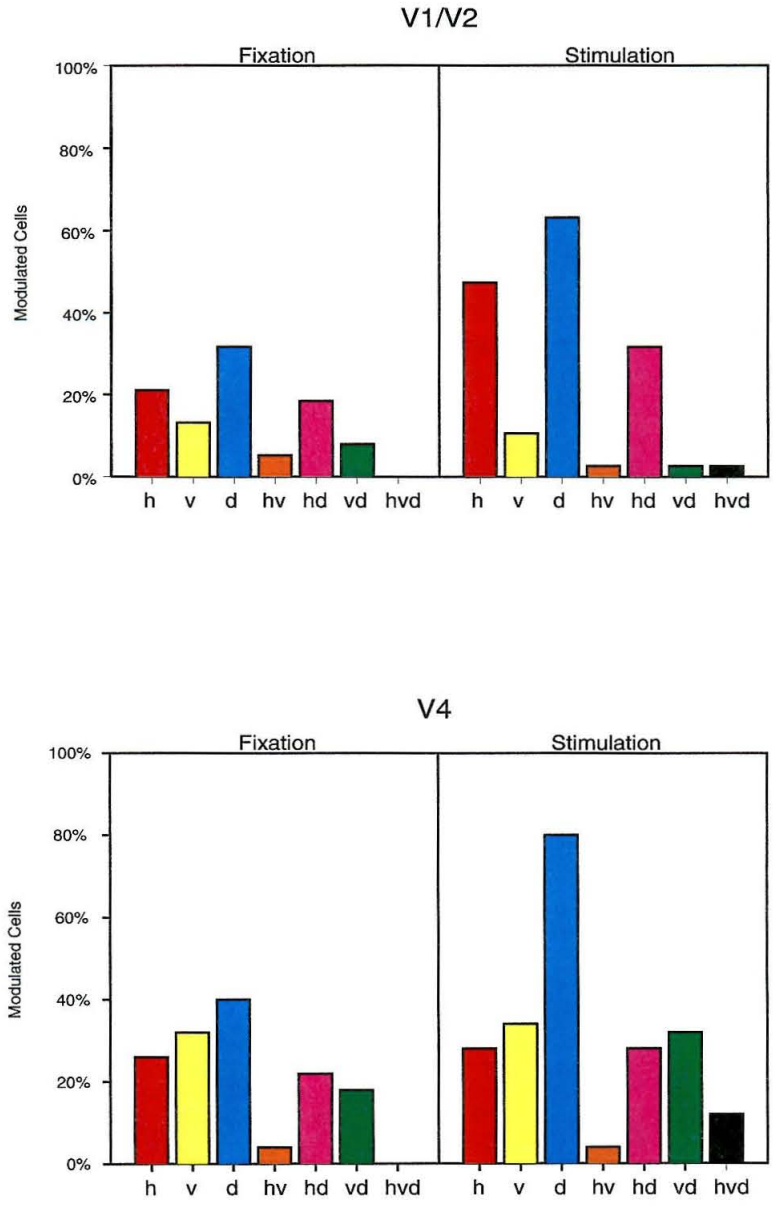
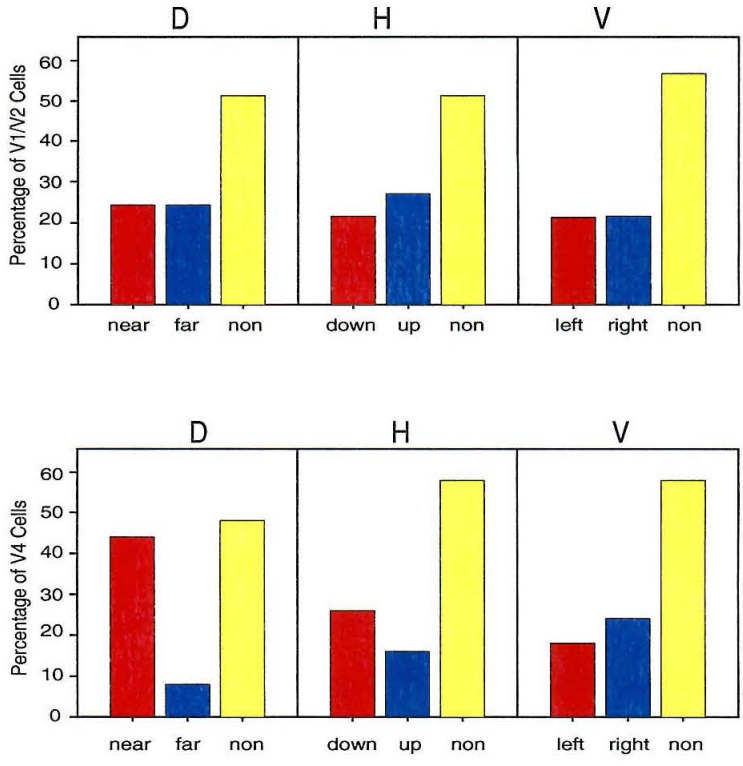


Figure 1.15: Three Way MANOVA
Summary results of three way MANOVA analysis for fixation only and stimulation periods. Percentage of cells that had significant ($p < .01$) modulation with respect to H,V,D, or some combination of these experimental variables. Results for V1/V2 and V4 are shown separately.



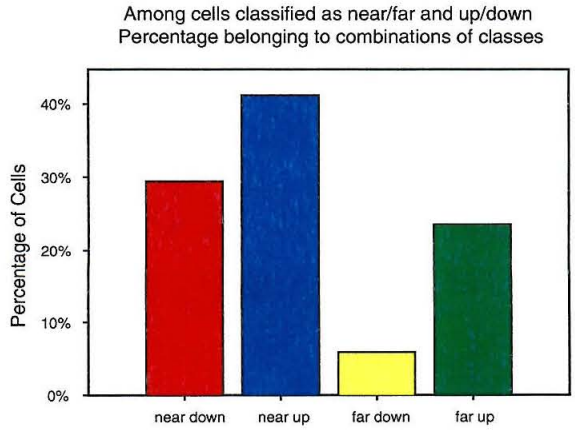


Figure 1.16: Modulation Classes
Distribution of different types of modulation.

of modulation with respect to visual cortical area is shown in figure 1.16 below.

In V1 approximately $\frac{1}{4}$ of the cells are monotone increasing, $\frac{1}{4}$ are monotone decreasing, and $\frac{1}{2}$ are non-monotone[14]. In V4, however, there is substantially larger proportion of nearness cells and a substantially smaller proportion of farness cells than are found in V1.

Since each cell is assigned three classifications, one for each experimental variable, these measures permit an examination of the size of the intersections of the different classes. The most notable observation to come out of this analysis is the very small percentage of cells classified as both far and down compared with the other (d, v) categories. The implications of this observation are discussed in later sections.

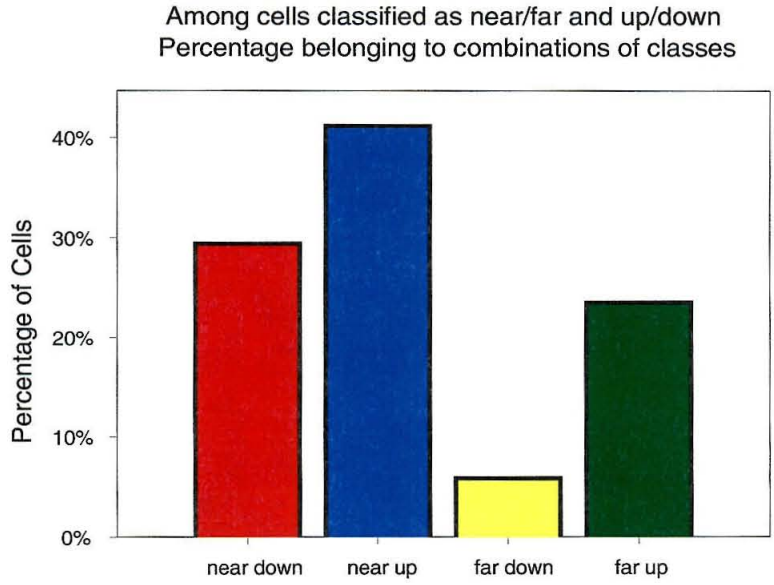


Figure 1.17: Distribution of (d, v) Classes

Percentage of population devoted to the conjunction of (d, v) classes consisting of near-down, near-up, far-down, and far-up.

1.4.2 Strength of Modulation

The degree to which a cell is modulated with respect to each of the experimental variables can be quantified using the fractional gain measures (Mod_h, Mod_v, Mod_d) described in the section on Data Analysis. These measures tell us on average the maximum amount by which

[14]The non-monotone half of the population is divided into $\frac{1}{4}$ concave and $\frac{1}{4}$ *convex, so that roughly* $\frac{1}{4}$ of the population falls into each of the four categories. Since we will not further discuss the concave/convex distinction, we have lumped them into a single non-monotone category.

the mean firing rate can change with respect to the experimental variable, normalized by the excitability of the cell. The graph below shows the distribution of modulation indices found in our population, which is roughly normal with a mean of 0.4.

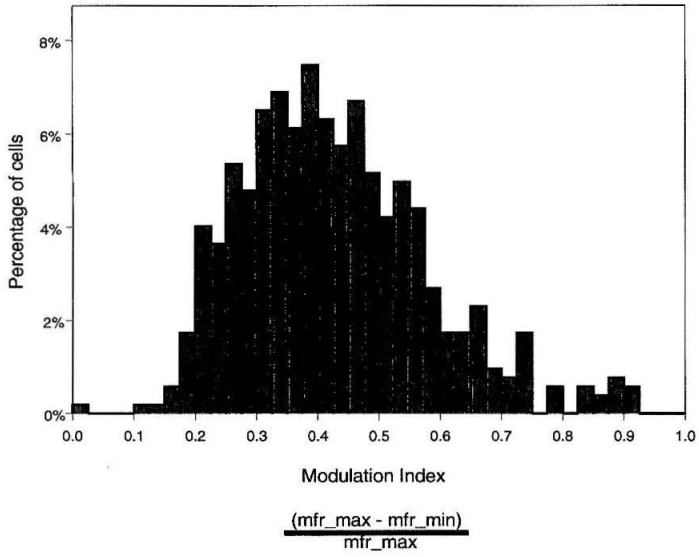


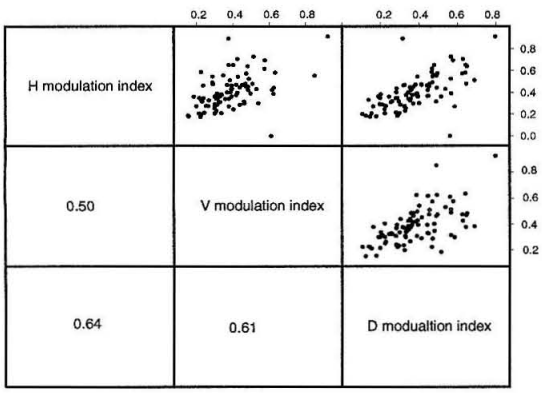
Figure 1.18: Histogram Showing Distribution of Modulation Indices for All Cells.

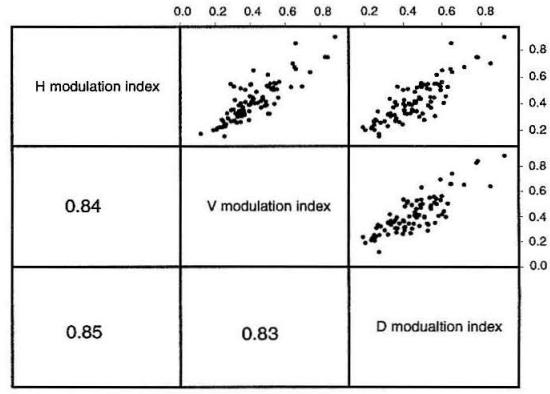
When separated out by modulation with respect to H,V, and D, fixation only modulation and stimulation modulation, and cortical area V1 and V4 the distributions were not significantly different from each other with the exception that modulation with respect to D was very slightly though significantly larger on average.

The analysis also revealed that there were significant correlations between the modulation indices (Mod_h, Mod_v, Mod_d) both during the fixation only and the stimulation period. The correlation was strongest between (Mod_h, Mod_v, Mod_d) during fixation only, and weakest between (Mod_h, Mod_v, Mod_d) for fixation only and (Mod_h, Mod_v, Mod_d) for stimulation.

1.4.3 Modulation in the Absence of Receptive Field Stimulation

Typically, in parietal visual areas where eye position modulation has been found, fixation activity also varies with gaze and appears primarily to convey information about eye position.[70] A similar modulation of fixation activity was found in V1, V2, and V4 [23]. In those experiments 47 percent of V1 cells and 51 percent of V4 cells showed significant





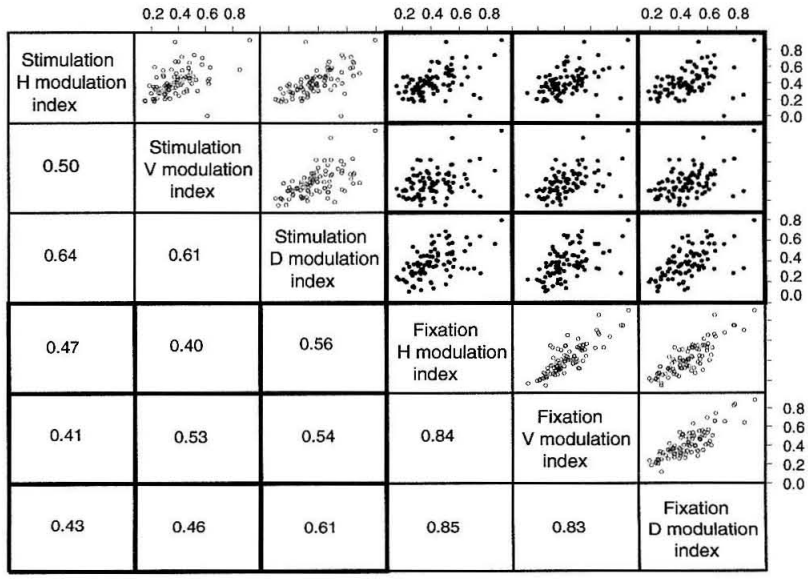


Figure 1.19: Left: scatter plot of modulation indices during stimulation period with correlation coefficients; Center: scatter plot of modulation indices during fixation only period with correlation coefficients; Right: scatter plot of modulation indices during both fixation and stimulation periods.

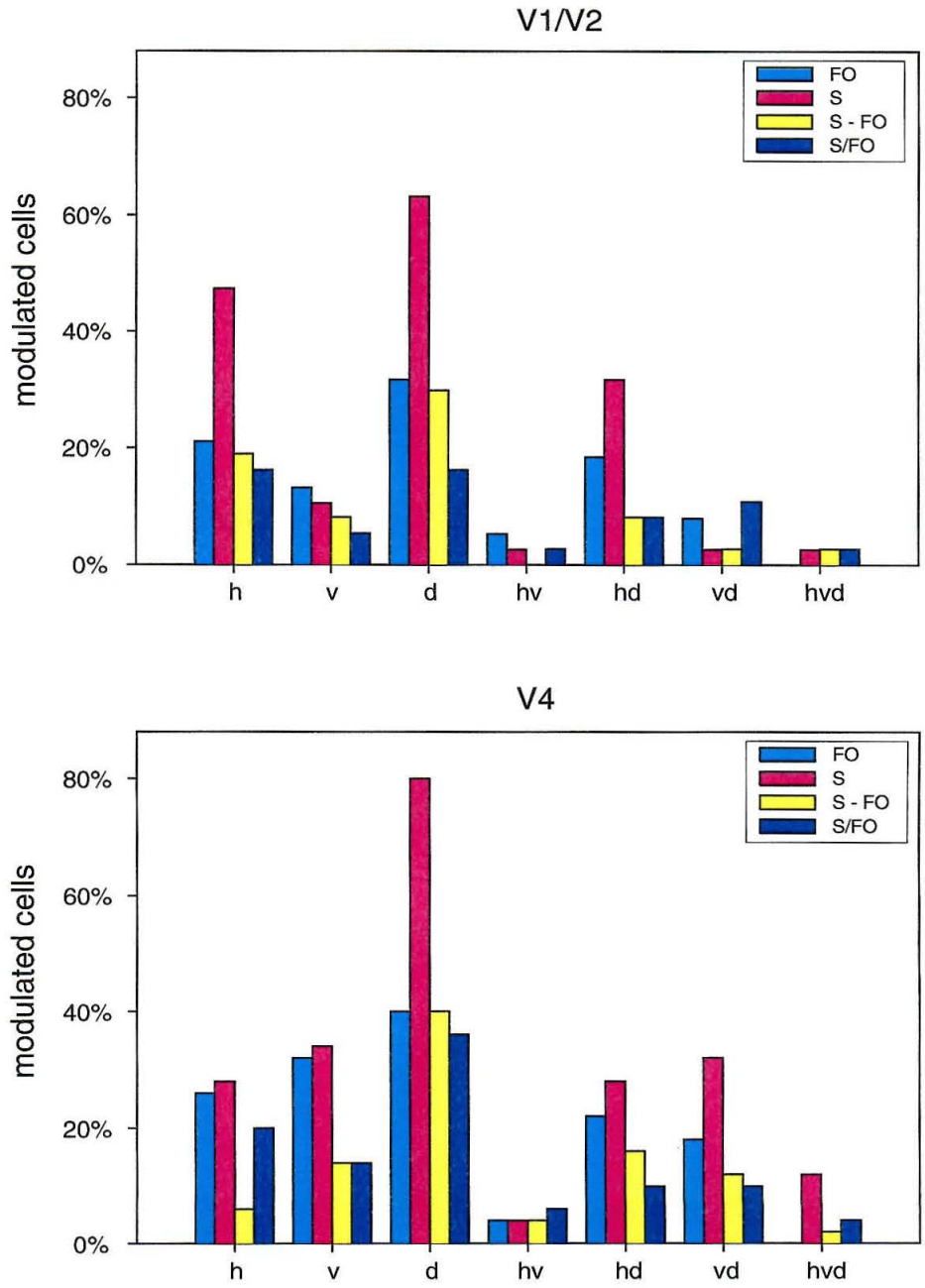


Figure 1.20: Three way manova: Fixation Only modulation factored out
Summary results of three way ANOVA analysis for fixation only period (FO), stimulation period (S), stimulation minus fixation (S - FO), and stimulation divided by fixation (S/FO). Percentage of cells that had significant ($p < .01$) modulation with respect to H,V,D, or some combination of these experimental variables. Results for V1/V2 and V4 are shown separately.

fixation only response modulation with respect to distance. In our experiments 40 percent of cells exhibited significant modulation during the fixation only period, corresponding to half of those which showed modulation during the stimulation period. Since the modulation effects for these two periods were not independent, we analysed the data to test for modulation effects remaining after factoring out fixation firing rate. Manova analysis was performed on data with fixation firing rate subtracted from stimulation firing rate, and for data with fixation firing rate divided out. The results are shown in figure 1.20. When fixation firing rate is factored out, by either subtraction or division, about half of the neurons still show significant modulation. This indicates that these results can only be partially explained by simple additive or multiplicative models of gain modulation by eye position.

1.4.4 Summary

While these experiments do not directly address the question of where the modulatory signals originate, the presence of modulation during the fixation only period strongly suggests that the modulatory signals are related to eye position. Tracing experiments have demonstrated an input to V2 and V4 from the small saccade part of frontal eye fields (sFEF)[77; 17]. The frontal eye fields (FEF) are an important component of the cerebro-ponto-cerebellar pathway involved in governing voluntary eye movements, including vergence and ocular accommodation[28]. There is a population of cells in FEF that display a tonic firing rate related to vergence angle and accommodation[28]. Stimulation of FEF and of V4 produce vergence and accommodation [42]. Thus the modulation seen in V2 and V4 may arise from efference copies of commands arising in the frontal eye fields. The modulation seen in V1 may result from indirect relay from from frontal eye fields via V2.

The lack of cells showing modulation with respect to both H and V may be indicative of an independence of the sources of the signals producing the H and V modulations. The difference between V1 and V4 in the number of cells showing modulation with respect to V may be an indication that information regarding V is not available to V1 and only enters the visual processing stream at the later stage of V4. This is further evidence of the independence of the sources of information regarding H and V. In contrast the coupling between the H and D modulations found in V1 may be indicative of a common source for the modulation with respect to these two parameters. This signal may be of the type hypothesized by Hering[38], that there are separate premotor conjugate and vergence eye

movement command.

Our results may contribute to a better understanding of the functional differences between the ventral and dorsal pathways in the visual cortex of primates. A basic distinction in these pathways is between the ventral specialization for object identity and the dorsal specialization for manipulation of objects in visual space [34; 58]. This distinction probably arouse in the evolution of the extra striate visual areas because the more ventral path proceeds from the foveal visual field representation in V1, whereas the dorsal path lies adjacent to the lower visual field representation where the hands are located during the manipulation of objects [57; 63]. Location in visual space is crucial for the performance of both ventral and dorsal functions but in different ways. For example in V4, a main component of the ventral path, object distance probably contributes to the mechanism of size constancy [23], which is crucial in discriminating object identity. The ability to accurately judge the size of objects at a distance requires gradual learning of the relationship between retinal size, object distance, and object size during early childhood [7; 75; 4]. Lesion experiments producing deficits on size constancy tasks indicate that this learning may occur in V4 and its upstream target IT[40; 80]. Such learned associations between eye position signals and sensori-motor contexts would have significant adaptive value.

The presence of a modulatory eye position signals in visual cortex prior to visual stimulation makes it possible for them to function as conditioning stimuli. Retinal stimulus characteristics (US) produce sensory responses in visual cortical neurons (UR). Learning resulting from repeated pairing of eye position signals (CS) with retinal stimulus characteristics (US) would tend to result in the eye position signal potentiating those neurons sensitive to the stimulus characteristics (CR), prior to stimulus presentation [5], thus preparing visual processing for the expected stimulus. A functional linkage between point of regard (CR) and the responses of visual cortical neurons (UR) learned in this way could result in perceptual learning of systematic relationships between point of regard and statistical characteristics of the visual environment. While there are circumstances in which strong correspondences exist between eye position and stimulus characteristics, and in these circumstances the visual system is capable of adapting to the eye position signal alone [47], in natural behavior it is more likely that eye position signals are but one of an array of extra-retinal signals that, when taken together, are very informative about the current sensory and behavioral demands, and strongly predictive of future sensory inputs. This array

probably includes eye position related signals relayed from frontal eye fields, fear related signals from amygdala, and reward related dopaminergic signals which serve a critical role in learning [39; 3; 9; 5]. All three of these extra-retinal signals converge on layer 1 of V2 and V4 and the frontal eye field and dopaminergic inputs also converge on layers 5 and 6 of V2 and V4.

Devices for human use as simple as rear view mirrors or bifocals, and as complex as a virtual cockpit explicitly based on a "what-you-see-depends-on-where-you-look concept," create correspondences between point of regard and distinctive information sources, and may be implicitly exploiting the natural talent humans have at learning such associations.

1.5 Discussion

1.5.1 Adaptation to Visual Context Through Specialization of Visual Processing

It has been suggested that specialists live in an effectively simpler world, and that the simplification permits faster and more accurate information processing and consequently higher behavioral efficiency[12]. There are many different types of context which provide hints for animals that they are operating within a simplified environment to which they can adapt by specializing. These contexts can be broadly categorized as ecological context, motivational context (drives, reward/punishment schedule), and task context (recognition vs. discrimination).

The adaptive value of specialized abilities has been studied for the task of finding and selecting foods. In insects, specialists are better and faster decision makers than generalists, which translates into higher quality diets and higher offspring survival. In primates it has been shown that quality of diet is a good predictor of lifespan and reproductive success. Skill in specialized visuomotor behaviors related to foraging, some of which depend heavily upon subtle visual discriminations (of color for instance), would ultimately have a significant impact on quality of diet. The location of objects in the visual field provides important clues about their identity. Object distance together with its retinal subtense reveals the size of an animal and whether it is a possible food item or a potential predator. Some threatening animals, like raptors, tend to be located in the upper visual field while others, like snakes, tend to creep in the lower visual field. The experience with their probable location

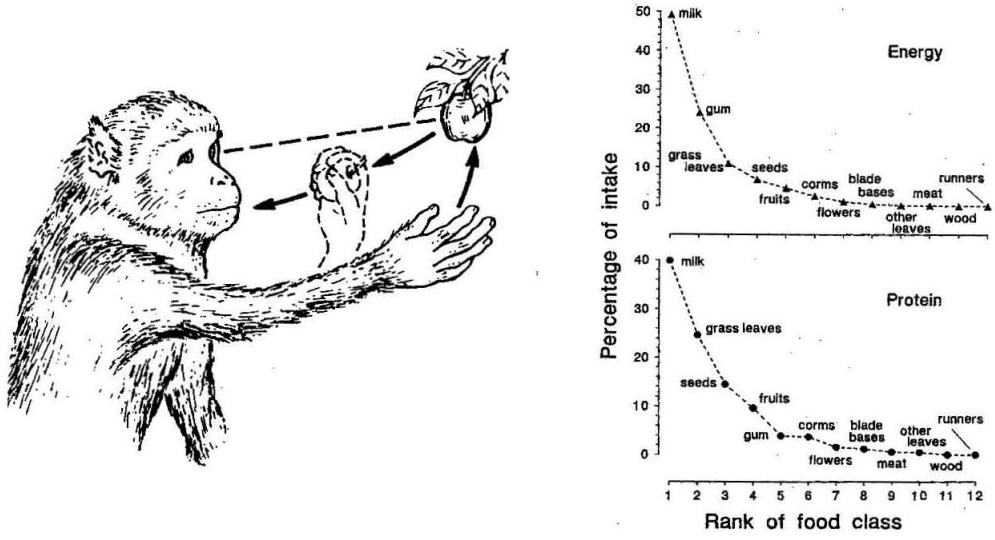


Figure 1.21: Eye Position Information and Visual Foraging
Drawing at left from[63, Previc]. Graphs at right showing the value of different food sources. Taken from [8, Altmann]

will facilitate their identification and speed the initiation of life-saving protective responses as illustrated in figure 1.22. Similarly, different types of food sources tend to be located in different parts of visual space, and this knowledge will facilitate efficient foraging[8]. Distance, in particular, may be a particularly strong cue for distinguishing between important classes of visual task and motivational contexts, such as between reaching distance (peripersonal) and walking distance (extra-personal) arenas. There is clinical evidence of cortical specialization along these lines[35].

In humans, there are a variety of different low level visual system adaptations specific to specialized visual tasks (ex.reading [64; 74; 30], driving a car [49]) which are able to coexist in visual cortex. Psychophysical studies of perception during reading reveal low level visual system specializations producing higher acuity for information presented to the right of the fixation point (in those that read from left to right)[64], contributing to the more efficient processing to written information. Other studies indicate that the minimum retinal size of legible letters is linked to distance cues such as the vergence angle[74]. Kohler's prism experiments also revealed interesting stimulus class specific or task specific types of adaptation. In inversion goggle experiments, subjects note that, after an adaptation period, "writing appears in the right place in the visual field and at first sight looks like

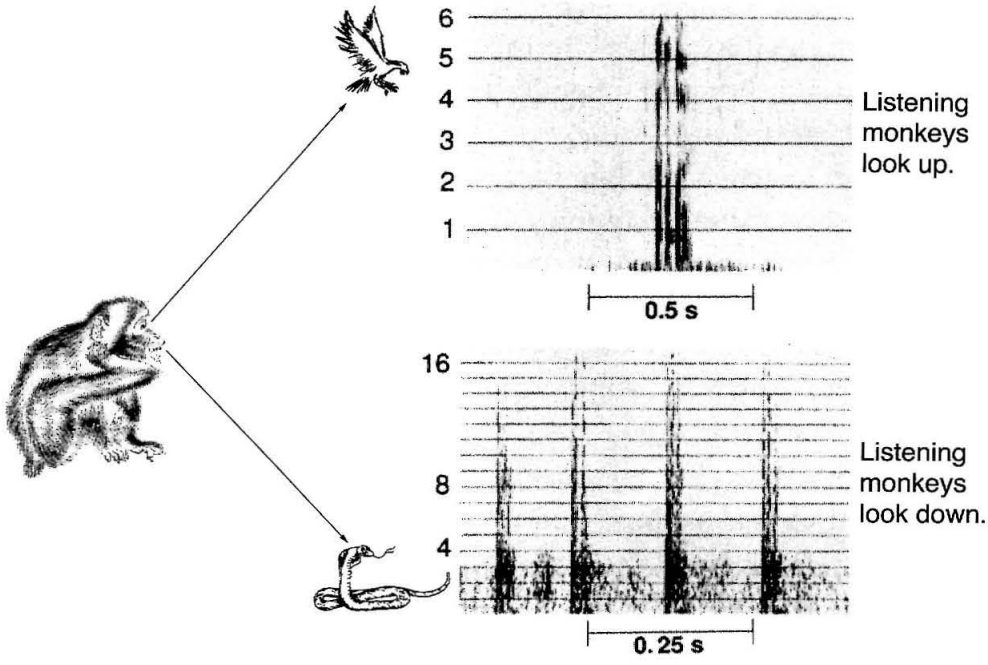


Figure 1.22: Warning Calls of Vervet Monkeys Correlated with Height in the Visual Field
Vervet monkeys have been shown to have different warning calls in response to danger from raptors and for snakes. The calls are illustrated with spectrograms. The protective value of these calls for the group comes from evoking joint attention to the danger. The correlation between the danger, the call, and the direction in the visual field is very important to this behavior.

normal writing, except that when one attempts to read, it is seen as inverted."[33, p206] In left-right reversal goggle experiments, subjects sometimes perceived that "a scene would come to look correct except that writing remained 'mirror-writing'."[33, p209] It would be interesting to determine if conversely it would be possible to adapt the visual system only in the reading task situation, without this adaptation influencing perception in other behavioral contexts (double dissociation of adaptation). Some subjects noted other shuffled or paradoxical perceptions which might be indicative of other types of specialized visual processing, for example "pedestrians were sometimes seen on the correct side of the street, when the images were right-left reversed, though their clothes were seen as the wrong way round!"[33, p209] The experiments of Ahissar and coworkers have found more generally that perceptual learning is specific to the stimuli used for training, and that the degree of specificity depends on the difficulty of the training conditions. "As *task difficulty* increases, learning becomes more specific with respect to both orientation and position, matching

the fine spatial retinotopy exhibited by lower areas. Consequently, we enjoy the benefits of learning generalization when possible, and of fine grain but specific training when necessary." [6]

Among the factors influencing the processing of visual information, the distribution of masking and attention over the visual field have been proposed important factors that become matched to the nature of perceptual and motor task being performed so as to amplify the information relevant to the task. "The events leading to visual awareness include a substantial editing process that de-emphasizes irrelevant information and adds interpretations and inferences about the meaning of the targeted information...this editing of visual signals begins in relatively early stages of processing in the cerebral cortex. What the observer is trying to see and what that observer knows about the visual scene have considerable impact on what is represented in the visual cortex. ..the cortex creates an edited representation of the visual world that is dynamically modified to suit the immediate goals of the viewer." [56] What is relevant to the task emerges from practice, and masking patterns develop in parallel with motor skill during the performance of a task [30]. In the task of reading, psychophysical evidence suggests that active lateral masking in the periphery effectively degrades background information into textural information, making it possible to process fine spatial detail at the center of vision without interference. This pattern of masking, learned early in life, is an important component of normal reading ability. Geiger et al. also suggest changing the distribution of masking over the visual field is but one of a lexicon of learned visual strategies, which are discrete (they do not shade into each other), and which one can switch between as appropriate for the task at hand.

1.5.2 Preparation in Sensori-motor Pathways

Adaptation of the visual systems of organisms to their environment occurs on many different time scales. On evolutionary time scales, the visual system of a species is adapted to suit the unchanging features of the environment in which they evolved. The spectral sensitivity of photoreceptor pigments of the retina and the shape and refractive index of the material forming the lens of the eye are good examples of such evolutionary adaptations[51]. There are many features of the environment which are not stable enough over time or consistent enough over the environment of the species to permit adaptation at the species level, but are stable enough over the lifetime of an individual to permit adaptation at the indi-

vidual level. Correspondingly, many species have the capacity during early development to learn general characteristics of their individual visual environment through alteration of anatomical connectivity during critical periods, or specific components of their environment through imprinting. The information acquired during these formative periods frequently remain stable throughout the lifetime of the individual. Years of training and gradual improvement during development are needed to acquire information about the highly complex but consistent correlations in the visual environment needed to perform specialized visual skills[15], such as face recognition and reading at mature performance levels[7; 75; 4]. On the time scale of months and weeks, prism experiments have demonstrated the flexibility of the visual system in adapting to dramatic distortions in the visual world[47]. Developmental changes, disease, trauma, fatigue, neuromuscular attrition, and environmental factors such as refractive correction, require a mechanism for continual recalibration and adaptation of the visual system on this time scale [16] [59; 46]. The same mechanisms which may have evolved for this purpose could be exploited for the development of task specific visual system specializations. Even relatively brief exposure to distinctive sensory and behavioral contexts during practice of novel visual tasks results in psychophysically measurable improvements in performance, phenomena termed perceptual learning and priming[6; 45; 21].

Preparatory activities which exploit the correlation between regularities in the visual environment and distinctive behavioral contexts result in faster reaction times, and more accurate or appropriate responses. Along the pathway from visual input to motor output there are many stages at which preparatory changes can influence visuo-motor behaviors. The term **preparatory set**, originally used to describe the pattern of activity found in primary motor cortex that reflected an animal's preparation to respond to a later stimulus[25], can be usefully applied throughout the sensori-motor pathway. On the sensory end, priming can be considered as an example of preparation in the sensory systems for expected incoming inputs. In visual cortical neurons, it is often difficult to separate the sensory response of a neuron attributable to the physical properties of the stimulus from memory responses resulting from the associative activation of the neuron as a member of an assem-

[15] By specialized we mean that these skills involve very narrow ensembles of images. Behaviorally they typically require discrimination of subtle visual features.

[16] Also relavent in this context is compensation in the nervous system, which entails the reassignment of functions in response to a change in the neural input or other signals caused by damage. The work of Mersnich, Ramachandran, and others illuminates the phenomena and the mechanisms behind reorganization of sensory maps in response to injury or sensory deprivation.

bly that encodes, by virtue of prior history, the experience to which the stimulus belongs[17] . This electrophysiological difficulty may be the source of the inseparability of memory from perception using behavioral or psychophysical techniques[27]. Perceptual memory and the process of perception are particularly closely tied when the memory recalled consists of values of physiological parameters that influence sensory processing, and when the recall of these parameters results in a change in physiological processing of information.

Small alterations in the processing at multiple stages along the sensori-motor pathway could potentially result in a cumulatively large alteration in the overall sensori-motor transformation. This would be a particularly useful strategy to use in systems where each particular stage has limited capacity to change[18]. Psychophysical experiments on perceptual learning phenomena suggest that learning most likely does occur at multiple levels of the visual system simultaneously, the strength of the learning in different areas being strongly dependent on the specific demands of the task. Evidence that alteration of sensori-motor transformations involves adaptation in multiple areas along the sensori-motor pathway dates as far back as the first inverted prism experiments of Stratton in the late 19th century. Stratton noted that he fairly quickly learned to overcome difficulties of performance while at that stage continuing to experience the scene as inverted. At a later stage the perception of the visual scene was upright. This observation, which has since been verified by many authors, highlights both the distinction between motor learning and perceptual adaptation, and the fact that both are components of sensori-motor adaptation[67].

The sensori-motor pathway is part of feedback loop which is closed through interaction with the environment: the motor system produces actions of the organism on its environment; actions on the environment change incoming sensory signals; the new sensory signals can lead to alteration of motor activity. Since the motor areas are the parts of the feedback loop closest to the sensory areas while still internal to the organism, the activity of areas on the motor output side of the sensori-motor pathway may be a particularly good source of preparatory cues to the areas closer to the sensory side. Von Holst and Mittelstaedt postulated a functional role for efference copy of motor commands such as these, in distinguishing

[17] "Strong recurrent synaptic connections in a neural network tend to produce stereotyped responses because population activity is primarily controlled by recurrent connections that do not depend on the stimulus. In such a network, the afferent, stimulus-dependent inputs serve to choose between a number of possible stereotyped responses."[71] Modulatory signals, including eye position signals, may also function to switch between stereotyped responses.

[18] The capacity of the visual system to adapt does degrade with age. Particularly striking are the differences in the plasticity before and after a critical period.

exafference (changes in sensory signals due to changes in the environment) from reafference (changes in sensory signals due to changes initiated by the organism). This function might be accomplished by formulating expectations of upcoming sensory events based on efference copy signals, and comparing these expectations directly with actual sensory data. While the types of expected sensory events envisioned by Von Holst and Mittelstaedt are those which are *causally* linked to motor commands, this idea can be extended to include expectation of sensory events which, while they have no direct *causal* relationship with the efference copy command, are nonetheless be *correlated* with the efference copy command.

Perceptual Learning

The presence of a modulatory eye position signals in visual cortex prior to visual stimulation makes them good candidates for conditioned stimuli in classical conditioning. When a sensory signal (the US), which produces an unconditional response in a given neuron (the UR), is consistently preceded by increased activity of a second neuron (the CS), the responses of the two neurons become increasingly correlated (CR). This increase in correlation represents the strengthening of the "functional connection" between the two neurons, and follows the Hebb-Stent Law[5]. [19] Likewise, we would expect the activation of a neuron by eye position signals during the fixation only period (CS), preceding the reception of retinal stimulation (US), would result in an increased correlation between this neuron and the neurons subsequently stimulated by the retinal input (CR/UR). After repeated pairing of eye position signals and visual stimulus, one would expect the eye position signal alone to potentiate the response of neurons sensitive to the expected visual stimulus (UR), prior to the presentation of the visual stimulus.

The idea of sensory conditioning has a long history dating back to the turn of the century. "A particular response tendency of a neuron can be referred to as a perceptual hypothesis ... Such an hypothesis may be set into operation by a need, by the requirements of learning a task, or by any internally or externally imposed demands on the organism. If a given perceptual hypothesis is rewarded ... it will become fixated; and the experimental literature ... indicates that the fixation of "sensory conditioning" is very resistant to extinction."

[19] Modifications were weaker when the stimuli that evoked the response carried no behavioral relevance. They concluded that the mechanisms of learning that underlie neuronal plasticity in the cortex of adult monkeys obey the essential features of both the Hebb-Stent Law and Thorndike's Law of Effect.

[15].[20]

Ivo Kohler, in his prism experiments during the 1960's, demonstrated the striking degree of flexibility the visual system has in adapting to distortions of the visual world. In the phenomena which he called **situational aftereffect** or **conditioned aftereffect**," a new perception is conditioned to the eye position stimulus. Thus, with prisms on, different retinal images ultimately come to signify the same phenomenal impression, depending upon eye position."[67, p204] In these experiments, "... there is a small distance between the eye and the prism. As a result the eye can, and frequently does, move with respect to the glasses...If one analyzes the geometry of the rays striking the retina, one finds that the adaptation problem is much more severe than if the prism and eye could be held in rigid relationship ... In fact, the distortion changes with every change in the angle that the axis of the eye makes in relation to the prism."[47, p433]

1.5.3 Code Switching

All of the adaptations referred to above are long lasting, and can be used effectively when the appropriate situation arises, even after long periods when they are unused. For example, in Kohler's goggle experiments, upon re-testing the subjects eight months later, they found that when the lenses were worn, the subject immediately showed the various modifications to behavior which had previously developed while wearing the spectacles. "It thus seemed that the learning consisted of a series of specific adaptations overlying the original perception, rather than a reorganization of the entire perceptual system."[33, p208]

The intermittent use of specialized adaptations of the visual system in behavioral or sensory contexts which arise discontinuously in the environment, requires a simple means by which special contexts can be recognized, and a means by which the appropriate specialized visual adaptations can be invoked. The learning and maintenance of multiple, specialized adaptations also requires a way of preventing catastrophic interference between the adaptations to distinct contexts, which otherwise might result in information acquired about one context overwriting previously acquired information about a different context[22].

There are two sorts of computational advantages of acquiring multiple specialized adaptations and switching between them as necessary. The first is more efficient use of expensive

[20] We will retain this view of a response tendency of a neuron as a perceptual hypothesis in the theoretical analysis since the terminology and concept map easily onto the concept of hypothesis in computational learning theory.

neural real-estate through time-sharing. When there are limitations on representational resources available for coding, extending the range and efficiency of representation, through the reuse representational elements for different purposes in different contexts, gives context sensitive languages great computational power[21]. The second advantage is improvement in behavioral performance. Dividing large problem domains into simple sub domains, and adapting to the different sub domains separately, permits faster convergence to more efficient processing behavior. As long as they are available inexpensively, hints providing information about the domain to which incoming information belongs can be exploited to increase coding efficiency by better matching coding characteristics to the characteristics of the information source.

The extent to which the visual system would be capable of switching operating modes to suit the needs of the current context depends in part on constraints regarding the cost of switching modes, the number of different modes that can be accommodated, and constraints on learning new contexts, all of which will be discussed in the section on theory.

Avoiding Catastrophic Interference

One solution to the problem of catastrophic interference is anatomical parcellation[22] where different types of stimuli are processed in physically separate cortical areas. The clinical evidence that damage to a localized visual cortical area can produce specific visual deficits, prosopagnosia being the best known example, provides evidence that such a strategy might be employed. More recent evidence, however, suggests that face responsive regions of visual cortex are used in many different types of tasks requiring visual expertise [37; 29]. The anatomical parcellation idea, when taken to the extreme, results in the grandmother cell problem. If the cortex needs a different area for the processing of each of the special environmental contexts which might arise, then there will be a combinatorial explosion in the number of areas needed. Hence anatomical parcellation is a very costly solution in terms of the amount of hardware needed, and cortex is metabolically very costly hardware indeed. Furthermore, psychophysical and electrophysiological evidence suggests that at least some components of context specific perceptual learning occur at very early stages of visual processing where the funnel for visual information is spatially narrow, and there is

[21] Since neurons are scarce in comparison with synapses, one way in which the same neurons might perform different functions in different contexts is by activating different sets of synapses in different contexts.

[22] Anatomical parcellation might more accurately be called spatial multiplexing.

little room for anatomical parcellation.

An alternative to the anatomical parcellation solution to preventing catastrophic interference is to use physiological parcellation [23]. In this strategy the same area or elements are used for different purposes at different times. There are many well known examples in neuroscience of systems in which neuro-modulatory signals result in switching of behavior in multifunctional neural circuits. The changes in both integrative processing of sensory signals and of ongoing motor output in response to modulatory signals [24] has been best illustrated and understood in lobster stomatogastric ganglion[54]. Such modulation of neural processing in response to signals indicating behavioral context may occur in the visual system as well.

While the goal of dividing the world into behaviorally distinct contexts provides clues as to how the world might be partitioned by a given species for use in code switching, conversely, the way in which an organism divides up the environment into distinct contexts may have an important impact on behavior. Acquiring information about a specific context has a cost, and to fully exploit this information behavior must be geared towards using existing learned contexts, rather than continually learning a new context for each new case[11]. One way of avoiding catastrophic interference is by restricting behavior so that situations always cleanly fall into one of the existing contexts, effectivly behaviorally filtering out situations which might fall into grey areas between contexts.

[23] More accurately referred to as temporal multiplexing

[24] Both chemical and electrical

Chapter 2 Theoretical Considerations

> ...the choice of which [representation] to use is important and cannot be taken lightly. It determines what information is made explicit and hence what is pushed further into the background, and it has a far-reaching effect on the ease and difficulty with which operations may subsequently be carried out on the information.
> *Vision, David Marr*

2.1 Introduction

This section is an exploration of a theory regarding the functional role eye position signals might be playing in ventral visual cortex. There are as many different roles that can be proposed for these signals as there are visual tasks, and many of the roles that have been proposed for specific computations are plausible[1]. Rather than taking a specific computational task as our starting point and trying to determine the role of the eye position signal within that task, we ask "Given that eye position signals are present in visual cortex together with associative learning mechanisms, how will these signals come to be utilized?" In answering this question we rely upon the evolutionary principal that coding in nervous systems must be understood with reference to the environments to which they are adapted; and the ecological observation that the environment is divided into physically and behaviorally distinct niches to which organisms have evolved specialized adaptations. The field of natural image statistics made the leap of applying the evolutionary principal to the understanding of neural encoding by interpreting the physiological characteristics of neurons in the visual system in terms of the empirically measured statistical structure of the environment to which the visual system is adapted[26; 69]. While the basic ecological observation has equally important implications for the understanding of visual systems[51], the study of the ecology of vision has been focused primarily on early stages of vision,

[1]Since V1 cells are known to respond to a very narrow range of disparities, it has been suggested that stereopsis requires either many cells tuned to each disparity at each locus, or a mechanism by which disparity tuning of individual cells is dynamically adjusted[19]. Extra-ocular eye position signals in visual cortex could potentially play a role in such a dynamic disparity adjustment.

with relatively little investigation into the implications for the encoding of visual information at cortical stages of processing. From the standpoint of coding theory this ecological observation can be interpreted as evidence that natural environments can be decomposed into distinctive sub-environments which are different enough from each other that there are significant advantages to developing distinct codes for representing each of these distinct sub-environments. In this chapter we explore the implications of such a decomposition for the encoding of visual information.

David Marr observed that "The usefulness of a representation depends upon how well suited it is to the purpose for which it is used... because vision is used by different animals for such a wide variety of purposes, it is inconceivable that all seeing animals use the same representations; each can confidently be expected to use one or more representations that are nicely tailored to the owner's purposes"[55]. Some specialist species are committed to very restrictive micro-environments and hence may require only one or few specialized codes designed to represent the limited variety of circumstances that may arise. Most species, however, live in environments containing many of these micro-environments. In this situation, a **code switching** strategy, in which specialized codes can be dynamically adopted to suit the current micro-environment, would provide both the flexibility of generalists with the performance of specialists. [2] In situations where there are reliable environmental cues available indicating a change from one distinct visual environment to another, many species have evolved the capacity to change operating modes of their visual systems depending on these cues. Switching between photopic and scotopic vision is a simple example of this strategy. A physiological mechanism permitting such changes to occur at the level of cortical processing would enable the visual system to adapt to very complex and transient partitions of the environment.

In previous sections we have discussed the survival advantages associated with task or context specific learning. From the computational standpoint the advantages of code switching arise primarily from the exploitation of a hint about the structure of the environment,

[2] Adaptive coloration in animals provides a good illustration of dynamically adopting a specialized adaptation to suit the current micro-environment. While walking sticks have adopted a morphological adaptation which commits them to the tree sub-environment, rabbits commonly have a seasonal coat of fur to match the seasonal visual properties of their environment; chameleons can adapt their skin color on a shorter time scale and can match a wider variety of visual environments encountered within its habitat; cuttlefish are the most dynamic of all, capable of adapting their skin pattern and texture continuously as they move around within their environment. Such dynamic switching of behavior to suit the current situation is commonplace in the animal kingdom.

namely that it can be decomposed into distinctive sub-environments. In computational learning theory hints are defined as auxiliary information about the target function that can be used to guide the learning process, and they help usually by reducing the size of hypothesis space which the learning algorithm has to search through[1]. Hints that split the global environment into simpler micro-environments effectively splits the learning of a complex function into piecewise simple components. Restricting learning problems to simple sub environments permits the use of simpler models, and often leads to faster convergence, and improved performance. One of the most interesting questions from the theoretical perspective is: How large a set of special cases should be maintained? If this collection becomes large enough, other costs associated with the space taken by this information and the difficulty of retrieving information from a large set begin to counterbalance the advantages.

2.1.1 Background

The Rice Machine is perhaps the earliest example of an algorithm using a codebook switching strategy. It was developed at the NASA Jet Propulsion Lab to improve the efficiency of encoding information from multiple distinct sources[20] and was used for sending image data to earth from the Mars Voyager spacecraft. The Rice Machine is a simple two-stage code, in which the first stage describes which code will be used[3] and the second stage describes the data using the chosen code. The multiple codebook idea has since been used to extend the Shannon source coding theorem to nonergodic stationary sources by using an ergodic decomposition to interpret a nonergodic source as a composite of ergodic sources[4] [32]. While a universal code is in theory more complicated than an ordinary code, involving many codebooks and a mechanism for switching between them, in practice it can be more efficient since separate codebooks can be used for distinct short term behavior.

In any two-stage coding scheme, there is a tradeoff between the average lengths of the first and second stage descriptions. The longer the first stage description is permitted to be, the more subtly differentiated the available code books can be, and the more well matched the codebook can be to the statistics of the data to be encoded. On the other hand, the

[3]One of four in the case of voyager, requiring a two-bit prefix to describe it

[4]This type of decomposition may be particularly relevant to describing nonergodic natural environments which may be most simply described as mixtures of ergodic sub-environments. Breaking the environment down into ergodic or approximately ergodic components has the great advantage that within each component learning the statistical structure of the component is valid.

length of first stage descriptions can easily overwhelm any gains that might be made in the second stage of encoding. One of the most important considerations in balancing these factors is the frequency with which codebooks will be switched. While the Rice Machine employed the simple scheme of re-evaluating which codebook to use at regular, and relatively short, intervals, more sophisticated schemes for determining when to switch codebooks have been developed. In particular, it is important to avoid a phenomena referred to in control theory as **chattering**, where switching between codebooks occurs too frequently, incurring very high overhead costs.

The mixtures of experts paradigm for supervised network learning is closely related to two stage codes. It was developed as a way of overcoming the drawbacks of training a single multilayer network to perform different subtasks on different occasions, most notably slow convergence rates and poor generalization due to catastrophic interference. "If we know in advance that a set of training cases may be naturally divided into subsets that correspond to distinct subtasks, interference can be reduced by using a system composed of several different expert networks plus a gating network that decides which of the experts should be used for each training case."[41] This technique is advantageous when the training set can be divided into simpler (homogeneous) subsets, and the learning task in each of these subsets is not as difficult as the original one. During training the gating network allocates a new example to few experts, and, if the output is incorrect, the corrective weight changes are localized to these experts. After training the network computes a value for an input by first having the gating input route the input to the appropriate expert and then having that expert compute the final output value. This scheme has been modified and extended in many ways since it was introduced, and the class of models it gave rise to are referred to as ensemble models or committee models. One notable extension of this model is the Boosted Mixture of Experts model, which replaces the gating network and the need to know in advance the appropriate partition of input space, with an algorithm which initializes a split of the training set to different experts and incrementally introduces new classifiers which are encouraged to become an expert on patterns on which the previous classifiers make errors or disagree[11].

The use of code switching strategies is becoming increasingly common in a wide variety of engineering applications. It is used in designing control systems[5] that can operate in

[5]They are often referred to as Hybrid Systems in the field of Control Theory[62].

multiple environments governed by distinct sets of equations. Transitions between these environments can cause the input-output characteristics of the control system to change rapidly or even discontinuously. Multiple models are needed in this case both to identify different environments and to control them rapidly[61]. Code switching is used for speeding CPU performance where detecting and predicting switches of computational context is critical to operating performance [53]; in text compression where model based encoding schemes arrive at different codes for different document classes[85]; and in the design of context aware networking environments where the system trys to detect the current context and anticipate the users needs.

2.2 Code Switching

Code Switching is most simply illustrated using the following modification of the classic Western Union Problem. The objective of Western Union is to maximize its profits by minimizing its costs assessed in terms of the number of characters that need to be sent over the wire. The typical strategy is to use a Huffman code which assigns small codeword to frequently occurring strings, and longer code words to less frequently used strings. The frequencies are typically estimated from large collections of messages collected over time. But now let us assume that Western Union has access to a simple and inexpensive piece of side information: each message is labeled with the name of the closest holiday at the time the message was sent, in the set of N holidays {fathers-day, mothers-day, xmas, graduation-time,...}. This extra information could be used to partition all past messages sent into N sets, estimate a separate probability distribution function for each element of the partition and construct an optimal codebook based on the different probability distribution functions. Now when the sender wants to send a message, he should first send a signal indicating which partition (or context) the message is coming from and hence which codebook to use to decode, and then use the associated codebook to code the message [6]. The likely outcome of using this code switching strategy is that "xmas" will have a very short encoding around Christmas time and a relatively long encoding around graduation time.

[6] In the case we are discussing, since both the sender and receiver presumably have access to a calendar and the current date, they could both know which codebook to use without any messages about context being sent. This is one advantage of making use of globally accessible side information.

2.2.1 Costs and Benefits of Code Switching

If one considers only the cost of the messages sent then clearly code switching saves Western Union some money. But one also needs to account for the the costs of computing N times as many frequencies, of computing N codebooks, of storing N codebooks, the cost of switching between codebooks, and the cost of sending a message indicating which codebook is being used. In addition, in order to compute the N sets of word frequencies, one may need N times as much data, which means that Western Union may have to wait a long time before they can capitalize on their database of messages. In this illustration, the side information was provided for free, but in many circumstances the side information also comes at a cost. When all of these costs are factored in, it is unclear whether code switching gives an advantage. Given a probability distribution over the environment and a specification of those costs mentioned above, we would like to determine the optimal partition of the environment, keeping in mind that one possible partition is no partition at all.

A code switching strategy is specified by a partition of the input space or environment along with function which classifies inputs into their partition sets; a codebook associated with each partition set; and an encoder which uses the correct codebook given the partition membership of the input. The costs associated with using code switching can therefore be broken down into the cost for obtaining a collection of codes, the cost for maintaining a collection of codes, a cost for using a code, and a cost for switching codes. The fact that codes and partitions must be learned, and that there are costs associated with learning, places constraints on the granularity and complexity of the partitions of the environment that can be practically used.

Using code switching requires the extraction two distinct types of information from the environment: the statistical regularities of specific contexts that can be exploited to produce an efficient codebooks; and the simple features of the context that can be used to identify it, to distinguish it from others,and, as a key, to recall the appropriate codebook. [7] [8] Both

[7]This may provide an interesting way of defining the distinction between procedural and declarative memory. Procedural memory has to do with the regularities found in a context. Declarative memory has to do with the identifying features of a context[22].

[8]The distinction between these two types of information is equivalent to the distinction drawn in linguistics and communication theory between types of information content referred to as transactional information and interactional information. Interactional signaling establishes the communications link and its characteristics prior to the sending of transactional information over the channel. In coding theory, a typical interactional signal might consist of the sender providing the appropriate codebook to the receiver before coded transmission begins. While there are systems that carry these two types of information in separate

types of information must be learned from experience and stored in memory for later use. Both the identifying features of a context and the contextual regularities may be obtained from within the same sensory modality. This is probably the case in many visual processes in which preattentive mechanisms provide the contextual cues[9], but the contextual cues can also come from a different sensory modality or even from signals internal to the animal (e.g., an internal clock).

The free availability of certain contextual cues may determine how the environment is partitioned by an animal using code switching. For this reason, use of the abundant internal signals as cues to context are of particular interest. The wiring devoted to connections between cortical areas accounts for a large fraction of the volume of the brain in primates. These connections have a high degree of macroscopical order, best described as a set of discrete bundles joining areas or parts of areas to each other, rather than as a continuum of intertwined fibers. There are wide variations in convergence and divergence of connections between cortical areas, but there is a tendancy for connections to be reciprocal. Among approximatly 30 visual areas in the monkey, there are 300 connections, about $\frac{1}{3}$ of all possible connections. Among 73 cortical areas in the monkey surveyed, about 15 percent of all possible connections exist. Connections between areas are more likely to exist when they are geometrically close to each other[13]. As mentioned earlier, a variety of extra-retinal signals are conveyed to visual cortex, including eye position signals. The specific set of extra-retinal signals present in visual cortex may be an adaptation that has evolved based on the value these signals have as side information about the environment.

2.2.2 Hints and Context: Learning Theory

We would like to consider when it is helpful for a learning algorithm to split the learning of a target function into a collection of learning problems where one is learning the target function on only a subset of the input space. It is similar to learning a complex function by first splitting it into piecewise simple components. The optimal way to accomplish this

channels, for example the telephone system, using what is called **out of band** signaling, many systems carry both in the same channel, using what is called **in band** signaling.

[9]These preattentive mechanisms and their role in directing eye movements may serve to direct the fovea of the retina to positions in the visual scene centered on distinct visual contexts. For example, if one of the contexts is faces, and there is a face in the visual scene, the fovea will be directed to the center of the face, rather than to the boarder between the face and the background or some other random position. Within the context of face discrimination, the fovea is directed to be centered on the eyes, nose, or mouth, and not other positions on the face (see the classic work of Yarbus [86]).

task depends on what functions can be represented simply by the hypothesis space, and also depends on what sort of simple functions can be produced by restricting the inputs to the target function f. Let us formalize the situation as follows: Let $(\vec{x}_1, \ldots, \vec{x}_n)$ be random variables drawn from sample spaces $(S_1, \ldots, S_n)$ which are a partition of the sample space S for the random variable $\vec{x}$. Each is drawn according to the probability distribution $F_{S_i}(x) = F(x|x \in S_i)$. Let $(\mathcal{B}_1, \ldots, \mathcal{B}_n)$ be subsets of hypothesis space $\mathcal{B}$. Now we have the following set of n learning problems: $(\vec{x}_i, F_{S_i}, \mathcal{B}_i, Select, Perf)$ where $i \in 1, \ldots, n$. Each learning problem selects its own hypothesis b_i after seeing k_i examples.

One way of solving any such learning problem is to take the k examples provided and partition the examples into subsets according to which of the sets $(S_1, \ldots, S_n)$ they fall into. Let us call these example partition sets, $(E_1, \ldots, E_n)$. Solve the i'th learning problem by selecting a hypothesis b_i on the basis of E_i. Finally, recombine these different hypotheses by defining b, the solution to the whole learning problem, as follows:

$$b(\vec{x}) = b_{h(\vec{x})}(\vec{x}) \text{ where } h(\vec{x}) = i \text{ when } \vec{x} \in S_i$$

The question is: "When does this strategy pay off?" Another way of phrasing the question is: **"Given a learning problem what is the optimal way to partition S and $\mathcal{B}$ for use with the above described learning scheme?"**, keeping in mind that one possibility is to leave S and $\mathcal{B}$ in one piece, which is equivalent to not using this strategy at all. Other, more restrictive and perhaps more realistic, versions of this problem might include either constraints on how S can be decomposed or on how $\mathcal{B}$ can be decomposed, or both.

The effect of such a hint is twofold:

- It restricts search of the hypothesis space to a smaller subspace $\mathcal{B}_i \subset \mathcal{B}$, thereby reducing the VC dimension and hence the number of samples needed to make a valid generalization.
- It reduces the number of samples by partitioning the input space.

This arrangement is only fruitful if there is enough of a reduction in VC dimension so that the reduced number of samples is still sufficient for valid generalization.

Let the VC dimension of $\mathcal{B}$ be $VC_{\mathcal{B}}$ and the VC dimensions of $\mathcal{B}_i$ be $VC_{\mathcal{B}_i}$. The fraction

of samples we can expect to get from each piece of the partition is $F(S_i) = p_i$. What we need to characterize is an expected generalization error. The hint we are given essentially amounts to our getting the function h in the formula for b for free. Characterizing how much effort it would have taken us to learn h might give us some insight into how much it can help us.

Another factor which might play an important role in determining the optimal partition of the input space is the effect of noise or the signal to noise ratio over the input space. Since the number of samples needed to make reliable estimates of statistical parameters increases with the magnitude of the noise, while the number of samples available will decrease as the partition of input space is made finer, there is probably a lower bound on the fineness to the partition determined by the noise in the input data.

Without placing some restrictions on how S and $\mathcal{B}$ can be decomposed, we might arrive at some absurd results. For instance, we can easily imagine a situation where the optimal partition of S is into sets on which f is constant, but these sets may have a very complex shape (i.e., the function h might have high complexity). Likewise, we can imagine the situation where the structure of the sets in the decomposition of $\mathcal{B}$ is highly complex. We need to be sure either to charge for the complexity of the decompositions, or at least set down some reasonable guidelines. Obviously, the nature of the optimal partition of S will have an important impact on the optimal partition of $\mathcal{B}$ and visa versa. The two extremes of the type of guideline that can be set down are the situation where the partition of S is fixed and no constraints are put on the partition of $\mathcal{B}$, and the situation where the partition of $\mathcal{B}$ is fixed and no constraints are put on the partition of S. I believe in most practical situations constraints must be placed on both.

The growth function which defines the VC dimension is defined as

$$m(N) = \max_{\vec{x}_1,\ldots,\vec{x}_N \in S} (\text{number of partitions of } \vec{x}_1, \ldots, \vec{x}_N \text{ by } b \in \mathcal{B})$$

Since we are now considering subsets of S and $\mathcal{B}$, we need to consider the VC dimension of a hypothesis set with respect to a particular input space, instead of assuming the full input space. The appropriate growth function is as follows:

$$m_{S_i,\mathcal{B}_i}(N) = \max_{\vec{x}_1,\ldots,\vec{x}_N \in S_i} (\text{number of partitions of } \vec{x}_1, \ldots, \vec{x}_N \text{ by } b \in \mathcal{B}_i)$$

In considering the relationship between the partition of S and $\mathcal{B}$, it is interesting to note that, just as hypotheses partition inputs, *inputs can be used to partition hypotheses.* On this basis we can define **the dual of the traditional growth function** as

$$\hat{m}(N) = \max_{b_1,\ldots,b_N \in \mathcal{B}} (\text{number of partitions of } b_1, \ldots, b_N \text{ by } x \in X)$$

and this can be used to define the *VC dimension of the input space with respect to a hypothesis set.* Using the dual growth function, it should be possible to derive results analogous to those expressing rates of convergence of learning in terms of the traditional growth function. These analogous results would tell us something about *the number of hypotheses we need in our hypothesis set for valid generalization*, rather than the number of examples we need for valid generalization. [10] In the typical learning problem, such a result would have little use since there are typically weak constraints on the number of hypotheses in the hypothesis set, and strong constraints on the number of examples available. But in the next section I will be discussing a natural situation where there is a severe constraint on the number of hypotheses in the hypothesis set.

These costs will be related to various observable properties of the code. For instance, we will make the reasonable assumption that the cost of maintaining a collection of contexts is a monotonically increasing function of the size of the collection. Hence we expect the size of the collection of contexts employed in a mode switching coding strategy is inversely proportional to the cost of maintaining this collection of contexts. In fact, one might be able to calculate the cost of adding a new context to a collection as a function of the complexity of the context to be added, and as a function of how similar the new context is to the contexts in the current collection. Also, the mean frequency of context shifts should be inversely proportional to the cost of switching contexts. It only pays to switch context if one is going to be in the new context for a significant period of time, otherwise the advantages gained by using a different context are outweighed by the costs of switching contexts.

The discussion of chapter 1 elaborated on the survival advantages of context specific sensori-motor adaptation and the evidence that the data on eye position modulation of responses of visual cortical neurons presented in chapter 1 might be interpreted as an instance of cortical adaptation of neural codes to the current context.

[10] These results would require defining a probability distribution on hypothesis space. We will give an example below of a natural definition of such a probability distribution function on a hypothesis space.

2.2.3 Natural Image Statistics

Research on natural image statistics largely neglects the fact that the image ensembles that the animal encounters are influenced not only by the visual environment but also by the behavior of the animal itself. Internal information about the state of the organism, such as eye position, could readily be exploited for hints about the specific context generating incoming sensory information or whether the current context has changed. Correlations which might exist and be exploitable in particular areas of the environment, are washed out in analyses of the ensemble of all images. Much greater advantage would be derived by partitioning the environment into simple sub-environments, each with their own distinctive and exploitable statistical structure.

It has been suggested that the receptive field properties of visual neurons are "matched" to the statistics of natural scenes. But when these researchers speak of "THE" statistics of natural scenes, they are trying to make a generalization about the ensemble of all natural images. [11] It would be of interest to analyze the statistics of several different ethologically relevant sub-ensembles of natural images and:

1. see if the statistics for these sub-ensembles are different (we will call them image contexts);

2. see if visual neurons can exploit information about image context and change their tuning properties appropriately.

Alternatively, by experimentally manipulating the reward associations, create arbitrary but easily identifiable ensembles of images which become behaviorally relevant.

In these experiments we postulate that the statistics on the ensemble of natural images can be parameterized in such a way that there is a systematic change in the statistics with respect to the parameter. It has been noted in the literature on natural image statistics that image statistics are invariant with respect to scaling (or change in focal length of the camera). Another simple parameterization of natural images is in terms of the focus of the camera. For these preliminary experiments we will only be concerned with two different focus distances: near, by which we mean within grasping distance; and far, by which we mean focus at infinity.

[11] Ironically, their empirical measurements are taken from a very restricted subset of this ensemble.

Chapter 3 Appendix: Relating Computational Learning Theory, Stock Portfolios, and Population Genetics

3.1 Introduction

These three seemingly unrelated topics each offers their own insights into the problem of optimally partitioning the environment. In computational learning theory, ensemble learning models use populations of learners with some variation in their abilities to improve performance over single learners. In population genetics, a population of organisms with some variation in phenotype has greater survivability than a homogeneous population, due to its ability to handle a greater degree of environmental fluctuation. In the branch of information theory called stock portfolio selection theory, the goal is to create an optimal population of stocks, with enough variability (diversified) so that it will not become extinct. There is a natural mapping between the problems encountered in population genetics, those in computational learning theory, and portfolio selection problems. We will describe the mapping between these three domains, and translate results derived in each of these fields into the language of the others. We focus on convergence theorems which are found in each of these fields, and how the concept of the VC dimension, which has been introduced relatively recently in computational learning theory as an important measure in estimating the convergence rate of learning algorithms, can be applied in both population genetics and portfolio selection problems. We will begin by describing the mathematical paradigms used in the portfolio selection problem, the computational learning problem of learning from examples, and the problems of predicting gene frequencies in a genetic population.

3.1.1 The Stock Portfolio Selection Problem

The stock market scenario we will be discussing can be described by the following tuple: $(\vec{x}, F, \mathcal{B}, Select, Perf)$ where

- $\vec{x} = (x_1, \ldots, x_m)$ is called the **stock market vector**. $\vec{x}$ is a random variable which

has probability distribution function F. Each x_i represents the *price relative* of stock i. The *price relative* is the ratio of the price at the end of the day to the price at the beginning of the day for stock i. There are a total of m different stocks. Each day $\vec{x}$ assumes a new value, and so the value of $\vec{x}$ on day j will be denoted $\vec{x}_j = (x_{j,1} \ldots x_{j,m})$. A record of N days of the stock market will be denoted $\vec{X}^n = (\vec{x}_1, \ldots, \vec{x}_n)$ so that $\vec{X}^n$ is an $n \times m$ matrix with $\vec{X}^n_{i,j} = x_{i,j}$.

- $\mathcal{B}$ is a set of permissible ways in which one can invest in the stock market.

$$\mathcal{B} = Simplex_m$$

where $Simplex_m$ is defined as

$$Simplex_m = \{\vec{b} = (b_1, \ldots, b_m) | b_i \geq 0, \sum_{i=1}^{m} b_i = 1\}$$

which is a simplex in m dimensions. [1] Each vector in this set is called a *portfolio* where b_i represents the fraction of one's wealth invested in stock i. At the end of each day, the investor is free to sell all his stock, and reinvest this money using a new portfolio chosen from $\mathcal{B}$, which may be chosen according the performance of portfolios on previous days. The portfolio chosen on day j will be denoted $\vec{b}_j = (b_{j,1}, \ldots, b_{j,m})$. A record of N days of investment portfolios will be denoted $\vec{B}^n = (\vec{b}_1, \ldots, \vec{b}_n)$ so that $\vec{B}^n$ is an $n \times m$ matrix with $\vec{B}^n_{i,j} = b_{i,j}$.

- $Perf$ is a performance measure which one wishes to maximize. Typically, one would like to maximize the expected *wealth relative* or the expected *growth rate*. The *wealth relative* on day j, denoted by S_j, is the ratio of the wealth at the end of the day to the wealth at the beginning of the day which is simply as follows:

$$S(\vec{b}_j, \vec{x}_j) = S_j = \vec{b}_j \cdot \vec{x}_j$$

[1] The basic model can be extended to account for side information by making the the elements $b \in \mathcal{B}$ functions which map values of the side information vector $\vec{y}$ to m-vectors on the simplex. This modification makes the set $\mathcal{B}$ more similar to a general hypothesis set as found in learning problems. We will expand on this later.

and the wealth accumulated over n days is:

$$S(\vec{B}^n, \vec{X}^n) = \prod_{i=1}^{n} S(\vec{x}_i, \vec{b}_i) = \prod_{i=1}^{n} \vec{b}_i \cdot \vec{x}_i$$

The *exponential growth rate* is the logarithmic counterpart to the *wealth relative* as follows:

$$W(\vec{b}_j, \vec{x}_j) = logS(\vec{b}_j, \vec{x}_j) = log(\vec{b}_j \cdot \vec{x}_j)$$

and:

$$W(\vec{B}^n, \vec{X}^n) = \frac{1}{n} logS(\vec{X}^n, \vec{B}^n) = \frac{1}{n} \sum_{i=1}^{n} logS(\vec{b}_i, \vec{x}_i)$$

One might want to place some additional constraints on the value to be optimized such as the permissible amount of variance in the value. In this way one can control the degree of conservatism in the investment strategy. [2] The performance measure can be a function of the history of the stock market and the previous investments made.

- *Select* is an algorithm which chooses a portfolio from $\mathcal{B}$ on the basis of previous experience and possibly some external information (insider tip?).

Once each of these elements is specified, the behavior of the system and its performance is completely determined (except for the behavior of the random variable which is not in our control). The concerns of research into portfolio theory applied to financial markets centers on finding an optimal strategy *Select* for choosing a portfolio when all other components of the tuple are specified.

3.1.2 Computational Learning Theory

The standard framework for formalizing the problem of learning from examples can be described by the same tuple $(\vec{x}, F, \mathcal{B}, Select, Perf)$ where:

- $\vec{x} = (x_1, \ldots, x_m)$ is called the **input vector**. $\vec{x}$ is a random variable which has probability distribution function F. Each x_i represents the value of variable i. Each iteration of the learning algorithm $\vec{x}$ assumes a new value, and so the value of $\vec{x}$

[2] It may also be prudent to take into account transition costs since some good strategies may incur very high transaction fees (brokers fees) due to the large number of transitions required by the strategy.

on iteration j will be denoted $\vec{x}_j = (x_{j,1} \ldots x_{j,m})$. A record of N samples from the input space will be denoted $\vec{X}^n = (\vec{x}_1, \ldots, \vec{x}_n)$ so that $\vec{X}^n$ is an $n \times m$ matrix with $\vec{X}^n_{i,j} = x_{i,j}$.

- $\mathcal{B}$ is called the *Hypothesis Set.* At each iteration one can choose a hypothesis from $\mathcal{B}$ on the basis of the performance on hypotheses chosen on previous iterations. A record of N iterations of hypothesis choices will be denoted $\vec{B}^n = (b_1, \ldots, b_n)$.

- $Perf$ is an error measure which one wishes to minimize also commonly referred to as the loss function and denoted $Q(\vec{b}, \vec{x})$. The goal in learning from examples is to find a hypothesis in $\mathcal{B}$ which is a good approximation for an unknown function f when all one is given is the values of f on randomly drawn inputs $\vec{x}$. The performance measure $Perf$ is a measure of how different the chosen hypothesis b is from the unknown function f. A typical example would be mean squared error:

$$Perf(\vec{b}, \vec{X}^n) = Q(\vec{b}, \vec{X}^n) = \frac{1}{N} \sum_{i=1}^{n} (f(\vec{x}_i) - b(\vec{x}_i))^2$$

Another important quantity is the *risk function* defined as

$$R(\vec{b}) = \mathcal{E}[Q(\vec{b}, \vec{x})]$$

where expectation is taken with respect to the distribution on $\vec{x}$. The *empirical risk function* is given by

$$R_{emp}(\vec{b}, \vec{X}^n) = \frac{1}{n} \sum_{i=1}^{n} Q(\vec{b}, \vec{x}_i)$$

Note that although the target function f is not explicitly included in the tuple describing the learning problem, it is implicitly present in the Performance function. While in Portfolio problems the goal is usually to maximize the performance measure since this corresponds to maximizing wealth, in learning problems the goal is typically to minimize the performance measure since it is a measure of error or loss.

- *Select* is a learning algorithm which chooses a hypothesis from $\mathcal{B}$ on the basis of previous experience and possibly some external information (hints).

There are a number of question of interest in this setting:

- How well does a particular learning algorithm (such as backprop) work in a particular context?

- Given a particular learning problem, what is the optimal learning algorithm to use?

3.1.3 Population Genetics

We can describe the situation in population genetics using the same tuple $(\vec{x}, F, \mathcal{B}, Select, Perf)$ where:

- $\vec{x} = (x_1, \ldots, x_m)$ where x_i represents the *fitness relative* of genotype i. The *fitness relative* in many cases is defined to be the ratio of the number of individuals of genotype i in generation j to the number in generation $j-1$ [3] but in artificial selection, the fitness can be judged by any criterion. There are a total of m different genotypes. $\vec{x}$ is a random variable which has probability distribution function F, which will be influenced by many factors including fluctuations in the environment. Each generation $\vec{x}$ assumes a new value, and so the value of $\vec{x}$ on generation j will be denoted $\vec{x}_j = (x_{j,1} \ldots, x_{j,m})$. Note that there is much debate about what constitutes fitness, but typically one might use number of viable offspring as a measure of fitness in which case the fitness relative is a measure of reproductive rate.

- $\mathcal{B} = Simplex_m$ is a set of possible gene frequencies in the population. [4] The elements of $\mathcal{B}$ are referred to as **genetic portfolios** and can be thought of as the number of individuals allocated to a particular genotype. The components b_i represent the proportion of genotype i in the population. The genetic portfolio generated on generation j will be denoted $\vec{b_j} = (b_{j,1}, \ldots, b_{j,m})$.

- $Perf$ is a performance measure. Many evolutionary biologists have proposed different

[3] Note that one interesting difference between the stock market and population genetics is that in the stock market the usual way to increase one's wealth is to have the price of individual shares rise. Occasionally, there will be situations where stocks split, in which case one gains not by having the price of stocks increase, but rather by having stocks "reproduce." In population genetics, the only way to increase wealth is by having shares (individuals) split (reproduce). One could imagine a stock market in which there were no prices, but the number of shares changed. The factor by which shares of a stock increased each day would be related to the profits of the corresponding company.

[4] The assumption, commonly made in population genetics, that purely random mating occurs, places an important constraint on the nature of $\mathcal{B}$. Given a population of a particular genetic composition and the assumption of random mating, only a limited class of gene frequencies can be generated. This constraint also limits the power of *Select*. For instance, it is apparent that changes in gene frequencies cannot change arbitrarily quickly.

performance measures which they hypothesize are being maximized or minimized by natural selection. Here I leave $Perf$ as simply an arbitrary function which may or may not be maximized by natural selection. A typical example of such a performance measure would be the expected *population fitness relative.* The *fitness relative*, denoted by the symbol S, is defined as the ratio of the population fitness in generation j the fitness at generation $j-1$:

$$S_j = \vec{b}_j \cdot \vec{x}_j$$

- *Select* represents natural selection as an algorithm which chooses a new distribution of gene frequencies from $\mathcal{B}$ on the basis of previous performance of the population. In contrast to the scenario in portfolio theory, the selection or investment strategy is fixed by nature (natural selection) [5]. In animal and plant breeding programs, we have the same situation as in the natural selection case except that *Select* is not fixed by nature. Breeders are free to choose at each generation the composition of the breeding population and the performance measure $Perf$, just as investors are free to reinvest at the end of the day and are free to choose the performance criterion $Perf$ on which to base their decision to reinvest. [6] The questions asked by animal and plant breeders are the same as those asked by financial investors: What is the optimal selection strategy *Select* to use so as to optimize $Perf$? They would like to choose the composition of the population in such a way as to maximize the wealth relative, which is a measure of the increase in performance of the population.

Most research in population biology centers on two questions:

- What exactly is the investment strategy embodied in natural selection? This question is approached by assuming a particular performance measure $Perf$ is being maximized. For example,

- What performance measures are being maximized? Or in other words what are the consequences of using the particular investment strategy employed in nature. This

[5] Although there will be some discussion later about mechanisms that have evolved for the purpose of altering this investment strategy by changing $\mathcal{B}$

[6] Artificial selection is not subject to the constraints of random mating. In fact, non-random, or directed mating, is probably the most important tool available to breeders for designing an investment strategy (that is until we can directly engineer the traits we desire).

question is approached by assuming a concrete implementation of the selection strategy embodied by natural selection. For example, the standard theory of stock market investment is based on the consideration of first and second moments of S, the objective being to maximize the expected value of S subject to a constraint on the variance. A similar objective may be at work in the evolution of organisms. It is desirable to maximize the fitness (or expected performance) of the population, but if the variance in performance is too great, extinction may result.

3.1.4 Stock Portfolios and Learning Theory

Portfolio Selection as a Learning Problem

Let us view the portfolio selection problem as a problem of learning from examples by simply equating the elements of the tuple specifying the portfolio problem $(\vec{x}, F, \mathcal{B}, Perf, Select)$ with the corresponding elements of the tuple for the learning problem $(\vec{x}', F', \mathcal{B}', Perf', Select')$. This mapping equates

- **stock market vectors** with **input vectors**.
- **portfolios** with **hypotheses**.

The only change we will need to make is in the definition of $Perf'$:

- $Perf'$ will be defined as follows:

$$Perf' = -log(Perf) = -logS(\vec{b}, \vec{x})$$

This makes the loss function for the learning problem map to the negative of the exponential growth rate. The following correspondences will then hold:

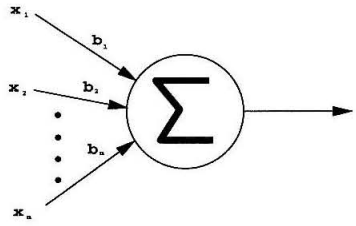
$$Q(\vec{b}, \vec{x}) = -W(\vec{b}, \vec{x})$$

$$Q(\vec{B}^n, \vec{X}^n) = -W(\vec{B}^n, \vec{X}^n)$$

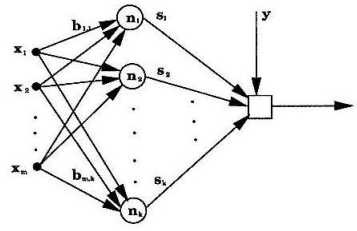
$$R(\vec{b}) = \mathcal{E}[Q(\vec{b}, \vec{x})] = -\mathcal{E}[W(\vec{b}, \vec{x})] \equiv -\bar{W}(\vec{b})$$

$$R_{emp}(\vec{b}, \vec{X}^n) = \frac{1}{n}\sum_{i=1}^{n} Q(\vec{b}, \vec{x}_i) = -W(\vec{b}, \vec{X}^n)$$

When this is done, it can be seen that the portfolio selection problem is equivalent to a rather simple learning problem. The hypothesis space for the learning problem consists of single node networks illustrated in the following diagram: where the weights of the network



are limited to the simplex $\mathcal{B} = Simplex_m$. As mentioned above in the description of learning problems, a target function is implicit in the definition of the performance function. If the performance function for the portfolio problem is taken to be the expected wealth relative, then we may ask 'What target function is implicit in the expected wealth relative?'. It will be argued here that this target function could reasonably be considered the Max function which returns the maximum value of all its inputs. The hypothesis set we are considering here is too weak to be able to approximate the Max function well when there is a uniform distribution on the inputs, but for some input distributions it may perform quite well. The hypothesis set can be strengthened by using side information. The use of side information described by Cover amounts to using a hypothesis set consisting of the following networks: The circle nodes in this diagram are the same summation nodes as appear in the single



node diagram. The output node in this case is a switch which can output any one of its inputs depending on the value of the side information $\vec{y}$. Obviously, this type of network can be significantly more powerful, depending on how large k is (the number of summation nodes), and what constraints are placed on the switching function. If no constraints are placed on the switching function, the overall function can be arbitrarily complex.

Using the tools of learning theory we can try to bound the number of examples (or days) needed to converge on a good solution in terms of properties of the hypothesis set, namely

the VC dimension. Later we will compute the VC dimensions of both the simple class of networks and the more complex class of networks.

Learning from Examples as a Portfolio Selection Problem

Let us now explore an alternative mapping between the portfolio problem and the learning problem. This mapping equates

- **stocks market vector** with **hypotheses performance vector**
- **portfolios** with **probability distributions on hypothesis space**

Since this mapping is not as straightforward as the one described above, let us make it more explicit. Given a learning problem described by the tuple $(\vec{x}, F, \mathcal{B}, Perf, Select)$ where $\mathcal{B} = \{h_1, \ldots, h_m\}$ and $|\vec{x}| = k$, we will construct a mapping onto a portfolio problem specified by tuple $(\vec{x}', F', \mathcal{B}', Perf', Select')$.

- $\vec{x}' = (x_1', \ldots, x_m')$ where

$$x_i' = \frac{e^{-Perf(h_i, \vec{x})}}{Cost(h_i)} = \frac{e^{-Q(h_i, \vec{x})}}{Cost(h_i)}$$

  represents the *performance relative* of hypothesis i. The fraction $\frac{e^{-Q(h_i,\vec{x})}}{Cost(h_i)}$ is the ratio of the dividends from the stock at the end of the day to the cost of the stock. In this market one rents stocks at the beginning of the day, and in return receives whatever dividends the stock generates during the day. The dividend of the stock is simply the performance of the hypothesis on the examples drawn during the current iteration of the learning process. At the end of the day the stock is returned[7], and one can take the money earned during the day and reinvest as one sees fit the next day. It is as if one is renting a piece of machinery (an algorithm) for the day, and whatever profit you can derive by using the machine is yours at the end of the day. We will initially assume that all stocks are the same price, which is a common situation in computational learning theory where each hypothesis in the hypothesis set has the same cost[8].

[7] This situation is actually not as strange as it sounds. It is realized in actual stock markets by ...

[8] But the cost of a hypothesis can usefully be employed to incorporate the complexity of the hypothesis into the choice of hypothesis. One would charge more for a more complex hypothesis, but would expect more dividends or higher performance from a more complex hypothesis.

The examples are drawn from the input space according to some unknown probability distribution function, the hypotheses are functions of this random variable, and the performance measure is a function of the input random variable and the value of the hypotheses on the random variable. Hence the vector of performance measures is a random variable with probability distribution function F'.

- $\mathcal{B}' = Simplex_m$ is a set of permissible ways in which one can invest in the hypotheses. Each *portfolio* in this set represents a probability distribution over the hypothesis set. The output of the portfolio selection algorithm will be a probability distribution on the hypothesis space rather than a specific hypothesis as is usually the case with learning algorithms.

- $Perf'$, the performance measure, is the same as in the Portfolio selection problem. Take it to be the exponential growth rate W,

$$W(\vec{b}_j, \vec{x}_j) = W_j = \log(\vec{b}_j \cdot \vec{x}_j)$$

This corresponds to the expected performance of the hypothesis set given a particular weighting function and a specific set of examples. The expectation is taken over the hypothesis set. The performance measure is a function of the input/output pairs generated by the hypothesis. Let us consider the special portfolio vectors $\vec{e}_i$ which have a 1 in position i and 0's elsewhere. The following correspondences hold:

$$W(\vec{e}_i, \vec{x}') = \log(\vec{e}_i \cdot \vec{x}') = \log(x_i') = \log(\frac{e^{-Q(h_i,\vec{x})}}{Cost(h_i)}) = -Q(h_i, \vec{x}) - \log(Cost(h_i))$$

If we set $Cost(h_i) = 1$ then we have

$$W(\vec{e}_i, \vec{x}') = -Q(h_i, \vec{x})$$

and

$$W(\vec{e}_i, \vec{X}'^n) \equiv W((\vec{e}_i, \ldots, \vec{e}_i), \vec{X}'^n) = -\frac{1}{n}\sum_{j=1}^{n} Q(\vec{h_i}, \vec{x}_i) = -R_{emp}(h_i, \vec{X}^n)$$

$$\bar{W}(\vec{e}_i) \equiv \mathcal{E}[W(\vec{e}_i, \vec{x})] = -\mathcal{E}[Q(h_i, \vec{x})] = -R(h_i)$$

- *Select* is a learning algorithm which adjusts the current probability distribution function over the hypothesis set on the basis of past performance. This is equivalent to choosing a portfolio from $\mathcal{B}$ on the basis of previous experience and possibly some external information (hints). One would like a learning algorithm to find a probability density function on the hypothesis space which maximizes the expected performance.

Before any examples have been generated, a learning algorithm will typically assume each hypothesis in the set is an equally good performer, so the initial portfolio will assign each hypothesis the same amount of wealth (or uniform probability distribution). As examples are generated the learning algorithm will modify the probability density function on the hypothesis space. If the target function is contained in the hypothesis set, then the probability density function should eventually converge to one of the basis vectors $\vec{e}_i$ where all the probability is concentrated on those functions in the set which represent the target function. *Under what circumstances will a probability distribution function over the hypothesis space perform better than any single hypothesis in the set?* This question will be the focus of later sections discussing **Context**. The answer to this question depends somewhat on how the probability distribution is used. There are a number of different ways in which this output can be utilized:

1. one can choose the hypothesis with the highest probability and use this one only.

2. each time a new input is generated, one can select a hypothesis from $\mathcal{B}$ according to the probability distribution function, and use the selected hypothesis on this trial.

3. each time a new input is generated, one can evaluate all hypotheses on this input, and produce the weighted average of these outputs (weighted according to the probability distribution) as the final output.

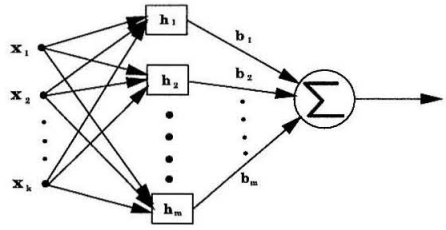
In the last case, the final result produced is *not* a hypothesis in $\mathcal{B}$, but rather a member of the hypothesis set $Hull(\mathcal{B})$ which is defined as follows:

$$Hull(\mathcal{B}) \equiv \{h = \vec{H} \cdot \vec{b} | \vec{b} \in Simplex_m\}$$

where

$$\vec{H} = (h_1, \ldots, h_m)$$

is the vector consisting of hypotheses in $\mathcal{B}$. The hypothesis set $Hull(\mathcal{B})$ can be seen as the set of networks which look like: Hence if we use the probability distribution in this way, then



the learning problem we have solved is described by the tuple $(\vec{x}, F, Hull(\mathcal{B}), \tilde{Perf}, Select)$ where:

$$\tilde{Perf}(h = \vec{H} \cdot \vec{b}, \vec{x}) = (Perf(h_1, \vec{x}), \ldots, Perf(h_m, \vec{x})) \cdot \vec{b}$$

Any learning problem in which the hypothesis set can be decomposed in this way into a linear combination of a finite number of hypotheses and where the performance function is linear in the performance of the hypotheses can be simply mapped to a portfolio problem on a hypothesis set of size m by considering each node in the layer preceding the output node to be a hypothesis. To intuitively understand the power of being able to combine multiple hypotheses in this way, let us briefly discuss how such an arrangement could potentially be exploited. Let's say one is faced with the problem of learning a function which is too complex to be well approximated by any single hypothesis set (in other words our hypothesis set is too weak). One could overcome this problem by just using the hypothesis set to approximate pieces of the target function (or more accurately approximate the target function when restricted to a small part of the input space). If one is provided with a good partition of the input space and side information telling which piece of the partition the input belongs to, then one can apply the appropriate hypothesis in the appropriate context and arrive at a good approximation of the target function where before there was none. This idea has many useful applications to design of learning algorithms, to design of portfolio selection algorithms, and to understanding the process of speciation in evolution.

Interesting questions can also be raised with regard to the reverse mapping, converting portfolio problems to learning problems. As we have seen, the simple way to do this is to equate portfolio vectors with hypotheses, but this section proposes an alternative, more complicated, but perhaps more interesting, mapping. When stocks are equated with

hypotheses through the mapping

$$x_i^{'} = \frac{e^{-Perf(h_i,\vec{x})}}{Cost(h_i)}$$

and one is given a stock portfolio problem, it is sensible to ask the following

- What do the hypotheses h_i correspond to?
- What does the target function correspond to?

If we consider the h_i's to correspond to companies, and $Cost(h_i)$ to be the price of company i's stock at the beginning of the day, then the function $e^{-Perf(h_i,\vec{x})}$ should give the price of company i's stock at the end of the day as a function of the company h_i and the environment $\vec{x}$. Such a function is obviously very complex. Included in the variables of the environment which affect the value of this function are current events, weather patterns, and the psychology of investors. But also included in these variables are factors which are more accessible and quantifiable, such as the profits of the company for the day, and structural characteristics of the company such as number of employees, number of levels of management, amount of feedback in the company organization, and the sophistication of their machinery. We might have some hope of characterizing the computational complexity of a company in terms of such structural characteristics.

We might also ask ' What is the target function implicit in the definition of $Perf'$? '. The optimal company.

3.2 Theorems on Convergence Rates

Portfolio theory, computational learning theory, and population genetics each have their own theorems regarding the number of days, iterations, or generations needed for a particular portfolio selection strategy, learning algorithm, or selection process, to converge to a stable and good solution. We will begin by stating a theorem about portfolios, introducing the appropriate terminology, and will follow this with theorems from computational learning theory, and population genetics, translated into the terminology from portfolio theory.

3.2.1 Comparison of Portfolio and Learning Theory Convergence Theorems

The best performance achievable in the portfolio selection problem described is the maximum wealth relative over all available sequences of investments:

$$S^*(\vec{X}^n) \equiv \max_{\vec{b}\in\mathcal{B}} S(\vec{b}, \vec{X}^n)$$

$$W^*(\vec{X}^n) = \frac{1}{n} log S^*(\vec{X}^n)$$

$$\bar{W}^*(\vec{X}^n) \equiv \max_{\vec{b}\in\mathcal{B}} \mathcal{E}[W(\vec{b}, \vec{x})]$$

We will refer to the portfolio vector which acheives this maximum as $\vec{b}^*$. A theorem of Cover and Ordentlich states that there is a portfolio selection algorithm called a μ weighted universal portfolio algorithm for which

$$W^*(\vec{X}^n) - \hat{W}(\vec{X}^n) \leq \log(C_{(m-1)}^{(n+m-1)})$$

where $\hat{W}$ is the wealth achieved by this portfolio selection algorithm, n is the number of days, and m is the number of stocks. The quantity $C_{(m-1)}^{(n+m-1)}$ is the number of ways of selecting a collection of n stocks when choosing from a set of m different stocks. This gives a bound on how different the actual wealth generated by the selection algorithm is from the maximal wealth that could have been generated, as a function of the number of days and characteristics of the set $\mathcal{B}$ of portfolios. We can also describe the quantities:

$$W'(\vec{X}^n) = \max_{\vec{e}_i} W(\vec{e}_i, \vec{X}^n)$$

$$\bar{W}' \equiv \max_{\vec{e}_i} \mathcal{E}[W(\vec{e}_i, \vec{x})]$$

These quantities maximize the wealth with respect to the portfolios at the vertices of the simplex. These portfolio strategies correspond to betting all of one's money on one stock. Note that $W' \leq W^*$ and $\bar{W}' \leq \bar{W}^*$.

Likewise in computational learning problems, the best performance achievable can be

characterized by the minima of the risk or loss functions:

$$R^*_{emp}(\vec{X}^n) \equiv \inf_{\vec{b}\in\mathcal{B}} R_{emp}(\vec{b}, \vec{X}^n) = R_{emp}(\vec{b}^*, \vec{X}^n)$$

$$R^* \equiv \inf_{\vec{b}\in\mathcal{B}} R(\vec{b})$$

The theorems of Vapnik and Chevonenkas state that for $C \leq Q(\vec{b}, \vec{x}) \leq D$, then

$$R(\vec{b}) - R_{emp}(\vec{b}, \vec{X}^n) \leq \frac{(D-C)}{2}\sqrt{\epsilon}$$

holds with probability at least $1 - \eta$, where

$$\epsilon = 4\frac{G^{\mathcal{B}}(2n) - \log(\eta/4)}{n}$$

and

$$R^*_{emp}(\vec{X}^n) - R^* \leq (D-C)\sqrt{\frac{-\log\eta}{2n}} + \frac{(D-C)}{2}\sqrt{\epsilon}$$

holds with probability at least $1 - 2\eta$.

If we have a mapping where:

$$R(\vec{b}) = -\bar{W}(\vec{b})$$

$$R_{emp}(\vec{b}, \vec{X}^n) = -W(\vec{b}, \vec{X}^n)$$

then

$$R^*_{emp}(\vec{X}^n) = \inf_{\vec{b}\in\mathcal{B}} R_{emp}(\vec{b}, \vec{X}^n) = \inf_{\vec{b}\in\mathcal{B}} -W(\vec{b}, \vec{X}^n) = -\sup_{\vec{b}\in\mathcal{B}} W(\vec{b}, \vec{X}^n) = -W^*(\vec{X}^n)$$

$$R^* = \inf_{\vec{b}\in\mathcal{B}} R(\vec{b}) = \inf_{\vec{b}\in\mathcal{B}} \mathcal{E}[-W(\vec{b}, \vec{X})] = -\sup_{\vec{b}\in\mathcal{B}} \mathcal{E}[W(\vec{b}, \vec{X})] = -\max_{\vec{b}\in\mathcal{B}} \mathcal{E}[W(\vec{b}, \vec{X})] \equiv -\bar{W}^*$$

The portfolio selection result

$$W^*(\vec{X}^n) - \hat{W}(\vec{X}^n) \leq \log(C^{(n+m-1)}_{(m-1)})$$

can be restated as follows:

$$\hat{R}(\vec{X}^n) - R^*(\vec{X}^n) \leq \log(C^{(n+m-1)}_{(m-1)})$$

The learning theory results:

$$W(\vec{b}, \vec{X}^n) - \bar{W}(\vec{b}) \leq \frac{(D-C)}{2}\sqrt{\epsilon}$$

$$\bar{W}^* - W^*(\vec{X}^n) \leq (D-C)\sqrt{\frac{-\log \eta}{2n}} + \frac{(D-C)}{2}\sqrt{\epsilon}$$

The mapping between portfolio selection problems and learning problems establishes a connection between the convergence rate of a learning algorithm and the doubling rate of a portfolio. In some situations (uniform fair odds) we know that a **Conservation Theorem** holds which stats that the sum of the doubling rate for the optimal portfolio and the entropy of the stock values is constant.[9] If the entropy of the stock market is high, then convergence will be slow. From computational learning theory we also know that the convergence rate should also be related to the VC dimension of the hypothesis set. Hence, it seems likely that VC dimension of the hypothesis set is folded into the entropy of the stock market. The higher the VC dimension, the slower the convergence and hence the smaller the doubling rate.

If we were dealing with a single deterministic hypothesis, the entropy of this function applied to the random variable generating inputs to the function would be bounded by the data processing inequality. This theorem states that the entropy of a deterministic function of a random variable must be less than the entropy of the random variable itself. But the population of hypotheses with the pdf defined on it constitutes a randomized algorithm. The entropy of a randomized algorithm applied to a random variable need not be lower than the entropy of the random variable. An additional entropy factor is contributed by the pdf over hypothesis space. Consider the simple situation where each hypothesis is a constant function, each outputting a different constant from the others. Since these functions disregard their inputs, the entropy of the input random variable makes no contribution to the entropy of the output. The entropy of the output in this case is simply the entropy of the pdf on the hypotheses. Note that if some of the hypotheses assume the same values, then the entropy would be lowered. Now let us consider another simple scenero in which there is only one hypothesis in the hypothesis set, and it maps inputs to outputs in a one to one fashion. In this case, the entropy of the output *is simply the entropy of*

[9] In less restrictive settings, an inequality holds.

the input random variable. Note that if the hypothesis were a many to one map then the entropy would be lowered.

Now let us modify this scenario. Based on the value of the input random variable, the previously constant outputs of the hypothesis will be permuted, so that a hypothesis which previously outputted value c will now output value d while another hypothesis will now output value c. In this situation the entropy can be either raised or lowered depending on whether the output distribution is made more or less uniform. A uniform distribution on inputs need not produce a uniform distribution on outputs. Conversely, a non-uniform distribution on inputs can produce a uniform distribution on outputs.

One factor which will certainly lower entropy of the output has to do with the characteristics of the hypothesis set alone. If the hypotheses are many to one maps and/or there is little diversity in the functions (i.e., it frequently happens that the outputs of two different functions are the same on the same input), then the entropy will decrease. This is related to the VC dimension of the hypothesis set.

3.2.2 Convergence Theorems in Population Genetics

In population genetics a specific hypothesis or function in the hypothesis set is described by its genetic code or genome. Individual genes are the components of the genome. In learning theory a hypothesis is described by a set of parameters, such as a set of weights in a neural network, and individual parameters are the individual components of this description. The majority of convergence theorems in population genetics are of the following form: If a particular value of a genetic parameter (allele) occurs in the population of size N with frequency p and it has a selective advantage s over its alleles in a randomly mating population, then the number of generations it takes for this value of the genetic parameter to become fixed is $G(N, p, s)$. The equivalent sort of theorem in computational learning theory would be: If a particular value of parameter value occurs with probability p in the hypothesis space, and the performance of functions having this parameter value is increased by a factor s, then the number of iterations of the learning procedure until the value of this parameter becomes fixed is $G(N, p, s)$. Typically, convergence theorems in learning theory regard convergence of all parameters simultaneously, and a probability distribution on the hypothesis space is usually not considered except perhaps for a uniform distribution. In the computational setting it would be extremely unusual to find a parameter value for

which the performance gain was constant and independent of the values of all the other parameters. In this respect the theorems of population genetics may not be particularly relevant to computational learning theory, but a more careful consideration of a probability distribution function on the hypothesis space and its impact on convergence, which is the meat of the theorems of population genetics, may well provide benefits in the computational setting. On the other hand, this assumption of the independence of the effects on fitness of changes in different genetic parameters might not be a good one even in population genetics (although it does make things more tractable). The theorems of computational learning theory regarding the convergence of all parameters, which do not assume independence, might be more applicable to natural populations.

3.3 Speciation as a Computational Strategy

The convergence theorems discussed thus far described the convergence of characteristics of the population (or equivalently hypothesis set pdf. or stock portfolio) to an equilibrium value. These theorems require some assumptions or ergodicity, or statistical uniformity, of the environment. For animal populations it is clear that the environment is not spatially uniform and this is why all species reside in a specific limited part of the environment. Within a limited domain the environment might be considered statistically uniform. Some species have limited their domain, or niche, to incredibly specific environments, while others, like humans, have domains which are quite large and diverse. In this section we will be concerned with the determinants of a species' range, and under what circumstances non-uniformity in the environment gives rise to the phenomena of speciation. We will view speciation as a computational strategy for dealing with non-uniform environments, and will examine how the concepts of hint, VC dimension, and over training, can help us understand the biological phenomena of speciation. In order to understand these phenomena, we must extend the ideas of convergence of properties of the population to convergence of both properties of the population and the domain which the population inhabits. There are in fact two sorts of evolution happening in parallel and interacting with each other: there is selection of genotypes by the environment; and there is selection of the environment by the genotypes. Finally, we will discuss how the computational strategy of speciation might be exploited in computational learning problems.

To illustrate the interaction of the partition of S and of $\mathcal{B}$ and how the dual VC dimension might be useful, I would like to discuss a natural application of these ideas to the theory of evolution. We will use the formalization discussed in section 1. [10]

From the standpoint of evolutionary biology, it is empirically apparent that the environment S is, in fact, partitioned into contexts, and these are usually referred to as niches. Operationally, niches can be defined either in terms of physical characteristics of the envi-

[10] A species is a collection of individuals that can interbreed to produce fertile offspring; in other words, the hypothesis set is a species if there is a function R which can take any two hypotheses from $\mathcal{B}$ and generate other hypotheses from them which also belong to the species. For any particular genotype, we can define a probability that the genotype will occur in the population, based on the existing population, the function R and the probability that two hypotheses will "mate". Two individuals/hypotheses are said to have distinct phenotypes if there exists a value of the environment on which the two generate distinct outputs or responses. Selection eliminates some individuals from $\mathcal{B}$ and applies R to others on the basis of their responses to the environment. This process generates a new hypothesis set from the existing hypothesis set.

ronment, or in terms of the spatial distribution of species. Frequently these two definitions of niche correspond. The most plausible explanation for the existence of this partition of the environment, and the partition of organisms into species which occupy these niches, is that such partitions are useful in learning or adaptation. By restricting the environment to which a species has to adapt, the species is able to more efficiently adapt to the environment.

There are two well known extremes to degree of niche specialization. Generalist species are wide ranging and able to adapt to a wide variety of environmental conditions. Specialist species can only survive in a narrow range of environmental conditions and are consequently highly localized. I hypothesize that this distinction can be stated simply in terms of VC dimension as follows: Generalist species have a high VC dimension while specialist species have a low VC dimension. A prediction of this hypothesis is that while generalists have the advantage of being able to adapt to a larger variety of environments, they will take longer to reach the same level of performance (or adapt to the same degree) than would a specialist species. Specialists have sacrificed the range of environments to which they can adapt, in favor of fast convergence to very high performance.

Two key questions in evolutionary biology are as follows:

- How is the partitioning of the environment into niches determined?
- Under what circumstances does speciation occur?

The interrelationship between the partitioning of the environment into niches and the partitioning of organisms into species illustrates the interdependence of optimal partitioning of input space and hypothesis space. From a consideration of this concrete situation, it is apparent that the partitioning of the environment found in nature depends not only on the characteristics of the environment itself, but also on the characteristics of the organisms which are adapting to the environment. [11]

[11] These questions are also related to issues regarding the causes of extinction and radiations of species. One of the proposed causes of extinction are rapid and dramatic changes in the environment. One way of protecting a species against extinction due to environmental fluctuations is to have phenotypic variation. The environment can be seen as assuming values which are used to evaluate individuals of a species and partition them into two groups: adaptive and maladaptive. Those in the adaptive group survive. A species will go extinct if the environment assumes a value for which all the individuals are maladaptive. If there is enough phenotypic variation, or in other words, if the dual VC dimension of the species is high enough, then no matter which value the environment assumes, there will be elements on both sides of the partition. As long as a few individuals (hypotheses) fall into the adaptive partition, the species will survive. Another way of protecting against the hazards of environmental fluctuations is to restrict the domain of the species to a sub environment where these fluctuations are smaller in magnitude or less frequent or both. These two strategies represent a tradeoff and a pair of competing forces on animal populations.

The former strategy dictates that every species should have a large and diverse population. Indeed I believe it might be possible to prove a "Nothing succeeds like success theorem" which states that the more populous a species, the more likely they are to survive in the future. A species might have a competitive advantage by virtue of its numbers alone. This fact would give rise to a pressure for species to expand their numbers. So why aren't all species numerous and diverse? The latter strategy dictates that every species should restrict its domain to the part of the environment which is most constant. This fact gives rise to the pressure for species to split into sub-populations (speciation), each of which is adapted to a small piece of the environment. So why aren't all species located in these environments? These two pressures, **the pressure to grow, and the pressure to split**, give rise to the partitions of organisms and environments that we observe.

Obviously the answer to both questions in the preceding paragraph lie in the fact that there are severe constraints on population size stemming from limited resources. Placing restrictions on the types of environment which are acceptable has the advantage of creating stable environments, but has the disadvantage of reducing the size of acceptable environment. The smaller the environment, the smaller the population size it can support, and consequently the smaller the amount of phenotypic variation. Given that there is a constraint on population size, one must decide the optimal way to distribute the population's phenotypic variation. The two extreme strategies are wide and sparse; and narrow and dense. These probably correspond to the strategies of generalists and specialists respectively. The narrow and dense strategy will do a better job approximating a function within a narrow domain than the wide and sparse strategy in the same narrow domain, but the wide and sparse strategy will permit the approximation of a wider variety of functions. [12]

[12] Adaptive radiations, the sporadic bursts of speciation found in natural history, can be understood in terms of the occasional discovery of new and very powerful hypothesis sets. A definition of evolutionary progress might be based on the idea that a more powerful hypothesis set requires fewer hypotheses for valid generalization than a weaker hypothesis set. Progress in evolution can be equated with the evolution of more powerful hypothesis sets.

Bibliography

[1] Y. S. Abu-Mostafa. Hints. *Neural Computation*, 7:639–671, 1995.

[2] R. Adolphs and D. Tranel. Preferences for visual stimuli following amygdala damage. *JOURNAL OF COGNITIVE NEUROSCIENCE*, 11(6):610–616, 1999.

[3] R. Adolphs, D. Tranel, and H. Damasio. Fear and the human amygdala. *Journal of Neuroscience*, 15:5879–5891, 1995.

[4] H. Philip Aeigler and H. Leibowitz. Apparent visual size as a function of distance for children and adults. pages 106–109, 1955.

[5] E. Ahissar, E. Vaadia, M. Ahissar, H. Bergman, A. Arieli, and M. Abeles. Dependence of cortical plasticity on correlated activity of single neurons and on behavioral context. *Science*, 257:1412–1415, 1992.

[6] M. Ahissar and S. Hochstein. Task difficulty and the specificity of perceptual learning. *Nature*, 1387:401–406, 1997.

[7] J. Allman. *Evolving Brains.* Scientific American Library, 1999.

[8] S. A. Altmann. *Foraging for Survival.* Chicago University Press, 1998.

[9] D.G. Amaral and J.L. Price. Amygdalo-cortical projections in the monkey (macaca fascicularis). *Journal of Comparative Neurology*, 230:465–496, 1984.

[10] R.A. Andersen. Coordinate transformations and motor planning in posterior parietal cortex. In M. Gazzaniga, editor, *The Cognitive Neurosciences.* The MIT Press, 1994.

[11] R. Avnimelech and N. Intrator. Boosted mixtures of experts:an ensemble learning scheme. *Neural Computation*, 11:483–497, 1999.

[12] E.A. Bernays and D.J. Funk. Specialists make faster decisions than generalists: experiments with aphids. *Proceedings of the Royal Society of London Series B-Biological*, 266:151–156, 1999.

[13] V. Braitenberg and A. Schuz. *Statistics and Geometry of Neuronal Connectivity.* Springer Verlag, 1998.

[14] B. Bridgeman. A review of the role of efference copy in sensory and oculomotor control systems. *Annals of Biomedical Engineering*, 23:409–422, 1995.

[15] J.S. Bruner and C.C. Goodman. Value and need as organizing factors in perception. *Jounal of Abnormal and Social Psychology*, 42:33–44, 1947.

[16] P. Buisseret and L. Maffei. Extraocular proprioceptive projections to the visual cortex. *Experimental Brain Research*, 28:421–425, 1977.

[17] J. Bullier, J.D. Schall, and A. Morel. Functional streams in occipito-frontal connections in the monkey. *Behavioural Brain Research*, 76(1-2):89–97, 1996.

[18] J.A. Buttner-Ennever and A.K.E. Horn. Anatomical substrates of oculomotor control. *Current Opinion in Neuobiology*, 7:872–879, 1997.

[19] V.A. Casagrande and J.D. Boyd. The neural architecture of binocular vision. *Eye*, 10:153–160, 1996.

[20] P.A. Chou, M. Effros, and R.M. Gray. A vector quantization approach to universal noiseless coding and quantization. *IEEE Transations on Information Theory*, 42(4):1109–1138, 1996.

[21] M.M. Chun and E.A. Phelps. Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience*, 2:844–847, 1999.

[22] N. J. Cohen and H. Eichenbaum. *Memory, Amnesia, and the Hippocampal System.* MIT Press, 1993.

[23] A.C. Dobbins, R.M. Jeo, J. Fiser, and J.M. Allman. Distance modulation of neural activity in the visual cortex. *Science*, 281:552–555, 1998.

[24] J.T. Enright. Perspective vergence: oculomotor responses to line drawings. *Vision Research*, 27:1513–1526, 1987.

[25] E.V. Evarts and J. Tanji. Reflex and intended responses in motor cortex pyramidal tract neurons of monkey. *J. Neuroscience*, 39:1069–1080, 1976.

[26] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, 1987.

[27] J.M. Fuster. *Memory in the Cerebral Cortex.* MIT Press, 1995.

[28] P.D.R. Gamlin, K. Yoon, and H. Zhang. The role of cerebro-ponto-cerebellar pathways in the control of vergence eye movements. *Eye*, 10:167–171, 1996.

[29] I. Gauthier. Expertise for cars and birds recruits areas involved in face recognition. *Nature Neuroscience*, 3:191–197, 2000.

[30] G. Geiger, J.Y. Lettvin, and O. Zegarramoran. Task-determined strategies of visual process. *Cognitive Brain Research*, 1:39–52, 1992.

[31] P. Gloor, A. Olivier, LF Quesney, F. Andermann, and S. Horowitz. The role of the limbic system in experiential phenomena of temporal lobe epilepsy. *Am. Neurol.*, 12:129–144, 1982.

[32] R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transations on Information Theory*, 44(6):1–63, 1998.

[33] R.L. Gregory. *Eye and Brain.* McGraw-Hill Book Company, 1981.

[34] C.G. Gross. Visual functions of inferotemporal cortex. In H. Autrum, R. Jung, W. Lowenstein, D. Mckay, and H.-L.Teuber, editors, *Handbook of Sensory Physiology Vol. VII/3B.* Springer-Verlag, 1973.

[35] P.W. Halligan and J.C. Marshall. Left neglect for near but not far space in man. *Nature*, 350:498–500, 1991.

[36] L.O. Harvey and H. Leibowitz. Size matching as a function of instructions in a naturalistic environment. *Journal of Experimental Psychology*, 74:378–382, 1967.

[37] J.V. Haxby, E.A. Hoffman, and M.I. Gobbini. The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4:223–233, 2000.

[38] E. Hering. *Spatial sense and movements of the eye.* Amer. Acad. Optom., 1942.

[39] J.R. Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309, 1998.

[40] N.K. Humphrey and L. Weiskrantz. Size constancy in monkeys with inferotemporal lesions. *Q. J. Exp. Psychol.*, 21:225–238, 1969.

[41] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[42] R.S. Jampel. Convergence, divergence, pupillary reactions, and accommodation of the eyes from faradic stimulation of the macaque brain. *Journal of Comparative Neurology*, 115:371–397, 1960.

[43] R.M. Jeo. Representation of three-dimensional space in primate visual cortex. *California Institute of Technology Thesis*, 1998.

[44] S.J. Judge. How is binocularity maintained during convergence and divergence. *Eye*, 10:172–176, 1996.

[45] A. Karni and D. Sagi. Where practice makes perfect: Evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences*, 88:4966–4970, 1991.

[46] W.M. King and W. Zhou. New ideas about binocular coordination of eye movements: Is there a chameleon in the primate family tree? *New Anatomist Review*, In Press.

[47] I. Kohler. Experiments with goggles. *Scientific American*, 206:62–72, 1962.

[48] R. Lal and M.J. Friedlander. Effect of passive eye position changes on retongeniculate transmission in the cat. *Journal of Neuophysiology*, 63:502–522, 1990.

[49] M.F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philisophical Transaction of the Royal Society of London B*, 352B:1231–1239, 1997.

[50] J. LeDoux. *The Emotional Brain.* Simon and Schuster, 1996.

[51] J.N. Lythgoe. *The Ecology of Vision.* Clarendon Press, 1979.

[52] J.C. Magee and E.P. Cook. Somatic epsp amplitude is independent of synapse location in hippocampal pyramidal neurons. *Nature Neuroscience*, 3:895–903, 2000.

[53] D. Mann. Speed system operation by matching cpu to need: understanding the many forms of context switching is key to maximizing risc performance in embedded-system applications. *Electronic Design*, 40:44–52, 1992.

[54] E. Marder. From biophysics to models of network function. *Annual Review of Neuroscience*, 21:25–45, 1998.

[55] D. Marr. *Vision.* W.H. Freeman and Co., 1982.

[56] J.H.R. Maunsell. The brain's visual world: Representation of visual targets in cerebral cortex. *Science*, 270:764–769, 1990.

[57] J.H.R. Maunsell and D.C. Van Essen. Topographic organization of the middle temporal visual area in the macaque monkey: Representational biases and the relationship to callosal connections and myeloarchitectonic boundaries. *The Journal of Comparative Neurology*, 266:535–555, 1987.

[58] A. David Milner and Melvyn A. Goodale. *The Visual Brain in Action.* Oxford University Press, 1996.

[59] M. Mon-Williams and J.R. Tresilian. A framework for considering the role of afference and efference in the control and perception of ocular position. *Biological Cybernetics*, 79:175–189, 1998.

[60] J.H. Morrison, P.R. Hof, and G.W. Huntley. Neurochemical organization of the primate visual cortex. In F.E. Bloom, A. Bjorklund, and T. Hokfelt, editors, *Handbook of Chemical Neuroanatomy, Vol 14: The Primate Nervous System, Part II.* Elsevier Science B.V., 1998.

[61] K.S. Narendra, J. Balakrishnan, and M.K. Ciliz. Adaptation and learning using multiple models, switching, and tuning. *IEEE Control Systems*, 5:37–51, 1995.

[62] A. Nerode and J.M. Davoren. Logics for hybrid systems. *IEEE Transations on Information Theory*, 44(6):1–63, 2000.

[63] F.H. Previc. Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, 13:519–575, 1990.

[64] K. Rayner, A.D. Well, A. Pollatsek, and J.H. Bertera. The availability of useful information to the right of fixation in reading. *Perception and Psychophysics*, 31:6:537–550, 1982.

[65] H.O. Richter, J.T. Lee, and J.V. Pardo. Neuroanatomical correlates of the near response: voluntary modulation of accommodation/vergence in the human visual system. *European Journal of Neuroscience*, 12:311–321, 2000.

[66] J.O. Robinson. *The Psychology of Visual Illusion.* Hutchinson University Library, 1972.

[67] Irvin Rock. *The Nature of Perceptual Adaptation.* Basic Books, Inc., 1966.

[68] J.P. Roll, J.P. Vedel, and R. Roll. Eye, head and skeletal muscle spindle feedback in the elaboration of body references. *Progress in Brain Research*, 80:113–123, 1989.

[69] D.L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1994.

[70] H. Sakata, H. Shibutani, and K. Kawano. Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey. *J. Neurophysiol.*, 43:1654–1672, 1980.

[71] E. Salinas and L.F. Abbott. A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Acadamy of Sciences*, 93:11956–11961, 1996.

[72] T. J. Sejnowski. The book of hebb. *Neuron*, 24:773–776, 1999.

[73] A.A. Sharp, M.B. Oneil, L.F. Abbott, and E. Marder. The dynamic clamp - artificial conductances in biological neurons. *Trends in Neuroscience*, 16:389–394, 1993.

[74] H. Shinoda and M. Ikeda. Visual acuity depends on perceived size. *Optical Review*, 5:1:65–68, 1998.

[75] A. Slater, A. Mattock, and E. Brown. Size constancy at birth: Newborn infants' responses to retinal and real size. *Journal of Experimental Child Psychology*, 49:314–322, 1990.

[76] R.R. Sokal and F.J. Rohlf. *Biometry.* W. H. Freeman and Company, 1995.

[77] B. Stanton, C. Bruce, and M. Goldberg. Topography of projections to posterior cortical areas from the macaque frontal eye fields. *Journal of Comparative Neurology*, 353:291–305, 1995.

[78] Y. Trotter and S. Celebrini. Gaze direction controls response gain in primary visual-cortex neurons. *Nature*, 398:239–242, 1999.

[79] Y. Trotter, S. Celebrini, B. Stricanne, S. Thorpe, and M. Imbert. Modulation of neural stereoscopic processing in primate area v1 by the viewing distance. *Science*, 257:1279–1281, 1992.

[80] L.G. Ungerleider, L. Ganz, and K.H. Pribram. Size constancy in rhesus monkeys: Effects of pulvinar, prestriate, and inferotemporal lesions. *Exp. Brain Res.*, 27:251–269, 1977.

[81] Leslie G. Ungerlieder, 2000. Personal Communication.

[82] J.L. Velay, R. Roll, G. Lennerstrand, and J.P. Roll. Eye proprioception and visual localization in humans: Influence of ocular dominance and visual context. *Vision Research*, 34:2169–2176, 1994.

[83] L. Weiskrantz. Behavioral changes associated with ablation of the amygdaliod complex in monkeys. *J. Comp. Physiol. Psychol.*, 49:381–391, 1956.

[84] Theodore G. Weyand and Joseph G. Malpeli. Responses of neurons in primary visual cortex are modulated by eye position. *Journal of Neuophysiology*, 69:2258–2260, 1993.

[85] Ian H. Witten, Alistar Moffett, and Timothy C. Bell. *Managing Gigabytes.* Morgan Kaufmann Publishing, 1999.

[86] A.L. Yarbus. *Eye Movements and Vision.* Plenum Press, 1967.

[87] H.Y. Zhang and P.D.R. Gamlin. The central thalamus of the primate: Neurons related to vergence and ocular accommodation. *Society for Neuroscience Abstracts*, 564.8, 1999.