THE UNIVERSITY OF CHICAGO

IS THE RETINA OPTIMIZED FOR PREDICTION?

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON COMPUTATIONAL NEUROSCIENCE

BY
JARED SALISBURY

CHICAGO, ILLINOIS
JUNE 2018

For my grandmothers.

Prediction is very difficult, especially about the future. — Niels Bohr

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisor, Stephanie Palmer, for her exceptional teaching, mentorship, and support; Olivier Marre for his mentorship and training in retina electrophysiology; Stéphane Deny for invaluable help in carrying out experiments; Heather Yee, Hanna Torrence, and Don Ho for collecting the natural movies; Leslie Osborne, Wei Wei, Nicolas Brunel, and Jason MacLean for serving on my committee and providing valuable feedback; all members of the Palmer Lab, past and present, for their ideas, conversations, and mathematical diversions; and my friends and family for their love and support.

ABSTRACT

In order to guide future behavior, nervous systems need to make predictions. This problem is so fundamental the we find evidence of predictive phenomena even in the retina. This thesis explores the hypothesis that prediction is a potential “design principle” for understanding the structure and function of the retina. We approach this hypothesis in several ways: First, we characterize statistically the motion of objects in a collection of natural movies, allowing us to quantify predictability in a natural setting. Then we test the predictive capabilities of the retina by recording population responses to artificial stimuli whose statistics are informed by natural object motion statistics; we find that responses are close to optimal when the stimulus statistics are in a naturalistic range. Finally, we examine neural responses to natural movie stimuli to determine if their structure is in line with our theory. A few mathematical derivations related to our theory are given in the appendices.

CHAPTER 1

INTRODUCTION

If the ultimate goal of neuroscience is to “reverse-engineer” the brain, then an important step will be to identify the design principles important to the “engineer.” This is a rather coarse analogy to the true situation: The brain’s engineer is the process of evolution itself, in which random variations, piled onto existing forms, persist according to the fitness they confer. We therefore expect the products of evolution to be idiosyncratic, reflecting their paths through the evolutionary tree, but also nearly optimal (or, at least, “good enough”) in terms of this abstract notion of fitness, through what amounts to a stochastic optimization process. The design principles in our analogy are educated guesses as to what constitutes fitness for a given biological system under study; a suitable theory can help us assess the validity of these principles by defining optimality in terms of experimentally measurable quantities. The following is an application of this approach to a particular candidate design principle—optimal prediction—in the context of the neural code of retinal ganglion cells.

Information theory provides a suitably general language for discussing neural design principles. First formulated by Shannon in the context of communication systems [49], it has proven to be an invaluable tool in neuroscience research [45], allowing us to quantify the amount of information (in bits), a sensory system conveys to the brain. Information is a statistical quantity that summarizes the dependance of two variables based on their joint probability distribution; it is therefore of the utmost importance to take these statistics into account when asking questions about coding.

A key insight is that information comes at a cost, namely, the metabolic energy required to maintain a population of neurons and to generate action potentials. This is particularly salient in the retina, where photoreceptors (the input cells) outnumber ganglion cells (the output cells) by a factor of ~ 100 in the periphery, reflecting the high metabolic cost of sending signals through the optic nerve to the central brain. Minimizing this metabolic cost while achieving a given rate of information (or, equivalently, maximizing information rate given a fixed amount of resources) forms the basis for the classic *efficient coding* hypothesis [6], which has proven remarkably successful in predicting the response properties of sensory neurons in the early visual system [50] and beyond [31] from first principles, using only knowledge about the statistics of natural stimuli.

Efficient coding is incomplete as a theory of sensory systems because it does not take into account the *relevance* of the incoming sensory information to the animal. More recent theoretical work [9] demonstrates how to take relevance into account using the information bottleneck method [53]. The information bottleneck method is a technique for (lossily) compressing incoming data as much as possible while retaining as much information as possible about a given relevance variable. In doing so it extracts only the relevant bits of information.

Many aspects of the environment are potentially relevant to the animal and could be incorporated into an information bottleneck framework. A particularly parsimonious choice of relevance comes from considering the fact that the world we live in is causal. By this we mean that sensory information is always *about the past*, yet its utility is always in guiding *future* behavior. For this reason we hypothesize that sensory systems selectively encode the predictive information in the stimulus [10], those features which allow us to predict the future stimulus. The past-future information bottleneck problem tells us how to navigate the tradeoff between efficiently encoding the past and maximizing information about the future. Importantly, it tells that an optimal solution must take into account not only the statistics of the past stimulus, but also the statistical linkage between past and future that makes prediction possible.

Most of our behaviors involve prediction in one form or another. Prediction is particularly important to sensorimotor behaviors because of the significant delays involved phototransduction, subsequent neural processing, and finally motor output. While prediction can, in principle, be performed at any stage of neural processing, compelling evidence from the retina literature suggests that prediction starts at the sensory periphery.

Far from acting as a simple signal transducer and pre-filter, the retina contains sophisticated circuitry to sift out behaviorally relevant information from the barrage of incoming photons [24]. The responses of the ganglion cells, the output cells of the retina whose axons form the optical nerve, are already highly processed reports of the input signals from the photoreceptor layer, and seem to reflect the complex structure of natural input—namely that it is composed of coherently moving objects. For example, one RGC subtype responds selectively to local motion but is suppressed by motion on a larger spatial scale, which separates the motion of objects from the self-generated motion caused by the animal's own movement [41, 4].

Even more strikingly, the retina seems to be able to *anticipate* the future position of moving objects. In [8], the authors demonstrate that flashing a bar of light onto a population of retinal ganglion cells produces a response with an appreciable delay of $\sim$ 50 ms, while moving the bar at a constant velocity to the same position leads to population responses that are spatially peaked at that position with zero delay. Thus, the *representation* of predictably moving objects in the retina appears to be compensated for the significant delay due to signal transduction. Similar results have been found in primary visual cortex [26, 51], and a perceptual analog called the flash-lag illusion has been studied psychophysically [37, 38]. When the moving bar violates its predictable motion by abruptly reversing directions, this triggers a rapid burst of spikes in the retina [48]; thus, the retina seems to convey both explicit predictions of the future and errors in those predictions. Other observations, such as the omitted stimulus response, in which some cells respond vigorously to a violation in a periodic pattern of stimulation [47], attest to the fact that the retina performs sophisticated computations in the context of prediction.

We would like to make clear the distinction between selectively encoding predictive information and *predictive coding*, which is a particular instantiation of efficient coding. Inspired by successful data compression algorithms from the signal processing literature [19, 20], predictive coding achieves an efficient representation of the stimulus by, at each time step, making a model-based prediction of the stimulus from its history, comparing this to the actual stimulus, and transmitting only the error, which will be small if the model is accurate; a downstream decoder can then reconstruct the stimulus from the stream of errors. In the most elaborate theories of predictive coding, the cortex is envisioned as a hierarchical Bayesian generative model, in which predictions are sent down the hierarchy while errors propagate upward [44, 7]. Despite its name, predictive coding does not actually solve the problem of finding an efficient representation of the future stimulus–it compresses the current stimulus (i.e., what is available to the encoder at a given time step) without regard to timing. For example, changing the signal transduction delay would lead to identical responses in a predictive coding framework, just shifted in time. In contrast, we argue that the retina must selectively encode the features of the past that are predictive of the future, and, in general, these optimal features change as a function of delay.

In this thesis we explore the problem of optimal prediction from several different angles. In Chapter 2, we introduce a method to extract the trajectories of moving objects from natural scenes, allowing us to empirically measure the statistics of natural motion which form the basis for prediction. In Chapter 3, we directly test the retina's ability to extract predictive information from a simplified moving bar stimulus, using two different statistical environments informed by these natural motion statistics. In Chapter 4, we address the more challenging problem of assessing optimal prediction in responses to complex natural movie stimuli, which requires a more indirect approach based in theoretical considerations. Finally, in Chapter 5, we draw these results together and make suggestions for future work.

CHAPTER 2

NATURAL OBJECT MOTION STATISTICS

2.1 Introduction

In order to make predictions in a visual world consisting of moving objects, we need to understand how objects move. At first this problem would appear to be solved by classical mechanics—given complete knowledge of all of the forces and masses involved in a given system, we can make extremely accurate predictions by applying the laws of physics. But the situation faced by animals in natural environments is rarely, if ever, so simple. It is impossible in practice to have complete knowledge of all of the relevant variables. Furthermore, we do not expect our system of interest, the retina, to contain elaborate physical models. Rather, it may take advantage of the statistical regularities caused by physical motion. Objects do not teleport; they move continuously from place to place, and due to their mass they exhibit inertia, the tendency to continue moving in the same direction (or remain stationary). We hypothesize that, over evolutionary time, the retina has evolved to take advantage of these kinds of regularities in order to efficiently encode information that is useful for prediction.

In this chapter, we present an approach to quantifying the motion contained in natural movies. The basis of our approach is a computer vision algorithm called the optical flow, which describes how one frame of a movie is transformed into the next by local translation of pixel intensities. Linking these translations through time results in trajectories that we can analyze statistically. These natural motion statistics are used in subsequent chapters to inform the construction of artificial stimuli and to help us interpret the results of experiments using natural movies as stimuli.

2.2 Results

We would like to build up a statistical description of how objects move. In particular, since objects move in a continuous fashion, we would like to characterize how correlated that motion is over time. To this end, we analyzed natural movies from the Chicago Motion Database, which consists of a variety of natural scenes collected for statistical analysis and for use as visual stimuli in experiments. All movies were recorded using a fixed camera, at high frame rates of 60 or 120 Hz, with subjects chosen to produce consistent motion within the field of view for minutes at a time. Subjects included flowing water, plants moving in the wind, and groups of animals like insects and fish (**Figure 2.1**). A forthcoming paper will contain a detailed description of the database.

2.2.1 Kurtotic motion distributions

The first step in our analysis was to compute the optical flow for all pairs of adjacent frames in our dataset using a standard optical flow algorithm [52] (**Figure 2.2**). The algorithm produces a continuous-valued estimate of horizontal and vertical image velocity (u, v) (in units of pixels/frame) at each pixel location (x, y). Each movie clip was 50 s long with a frame rate of 60 Hz (we subsampled movies filmed at 120 Hz), yielding $2,999$ optical flow frames per movie, or $47,984$ total. From these we estimated the probability distribution of object velocities by calculating histograms (using 0.25 pix/frame bins), shown in **Figure 2.3a**. Optical flow is measured in terms of pixels on our camera sensor (or degrees of visual angle) rather than objective physical units; an object moving at the same physical velocity viewed at two different depths will lead to different image velocities. Given that the image velocity is somewhat arbitrary, we scaled the velocity by its peak absolute value for each movie to facilitate comparisons across movies (**Figure 2.3b**). We performed the same analysis for the speed (taking the magnitude of each velocity vector) (**Figure 2.3b,c**).

Since all of the movies contain a significant amount of stationary background pixels, the

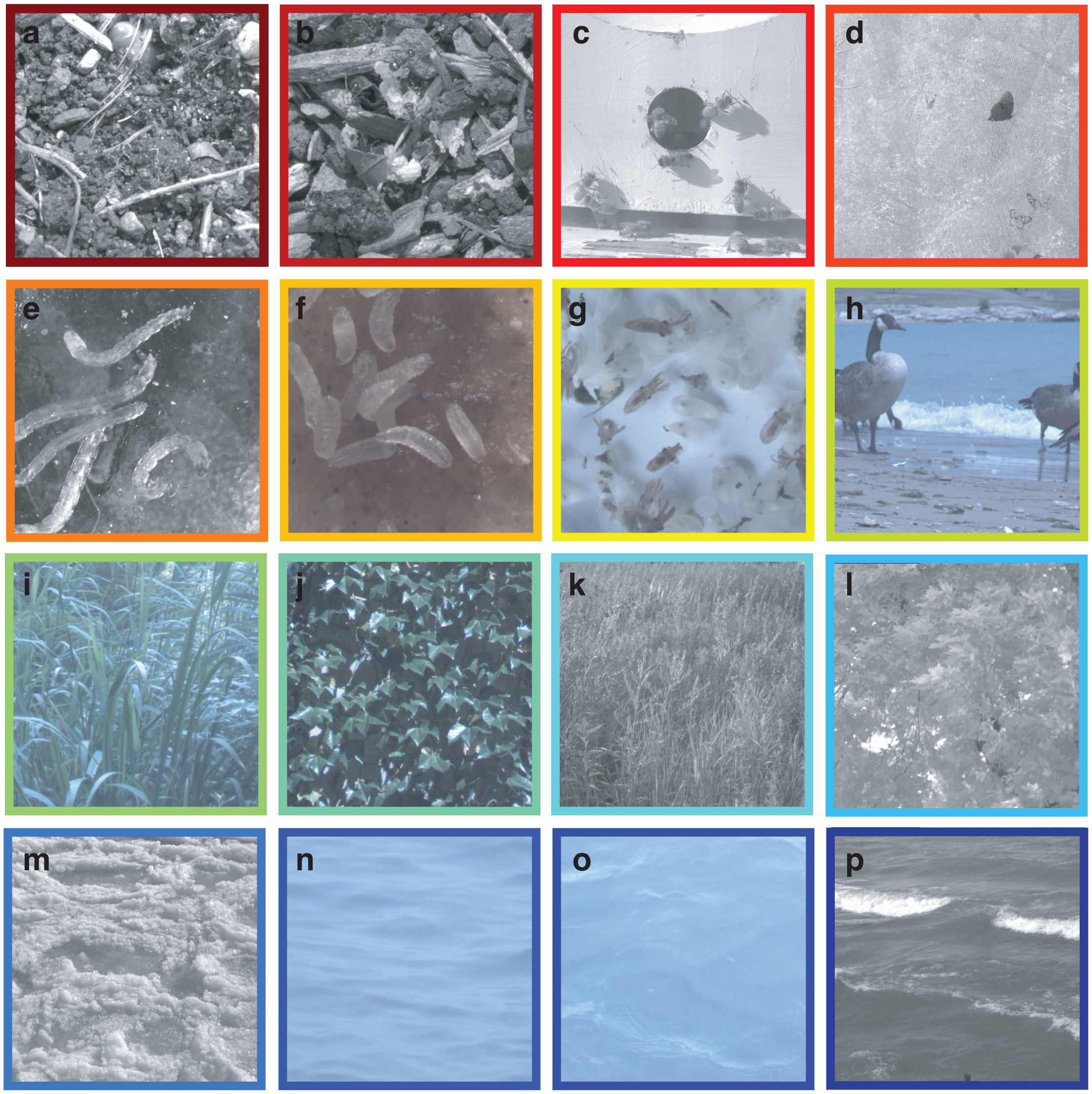

Figure 2.1: **Natural movie ensemble. a-p.** Example frames from 16 movies in the Chicago Motion Database. Border colors serve as a legend for subsequent figures. The movies depict (**a,b**) ants, (**c**) bees, (**d**) butterflies, (**e,f**) larvae, (**g**) baby octopuses, (**h**) geese, (**i,j,k,l**) plants blowing in wind, (**m**) floating ice, and (**n,o,p**) water.

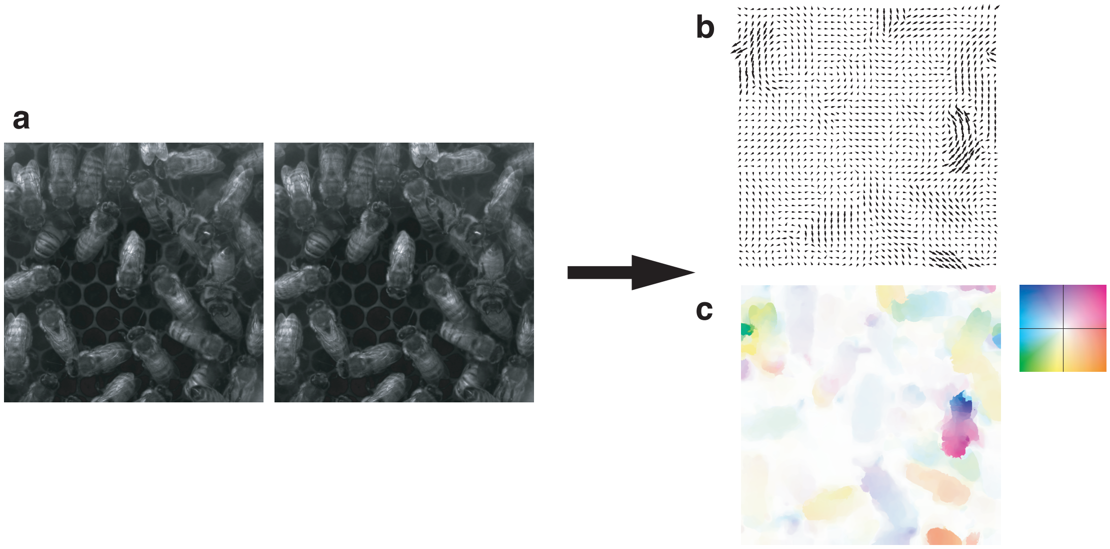

Figure 2.2: **Optical flow schematic. a.** Two subsequent frames from a natural movie in the Chicago Motion Database, depicting a bee hive with a glass cover. **b.** The optical flow for the two frames, depicted in vector form. Each arrow represents the optical flow vector in that location. Only a subset of pixel locations are used for display purposes. **c.** The same optical flow displayed using a colormap (right) in which hue corresponds to the angle of the optical flow vector at a given pixel location and saturation corresponds to magnitude (scaled by the maximum value). Both **b** and **c** were generated using the optical flow Matlab package described in [52].

probability distributions are heavily zero-inflated. They are best viewed on a logarithmic scale, which is an indication of kurtotic behavior. We find that velocity and speed distributions are nearly log-linear for a wide range of values, away from the inflated peak and noisy tails. Log-linearity is a signature of the exponential function, so we conclude that speed follows an approximately exponential distribution and velocity components follow an approximately Laplace distribution (two exponential functions decaying away from zero). These distributions tell us something fundamental about motion: Compared to the Gaussian distributions we often find in nature, motion distributions are *kurtotic*. This means that small and large values are more common while intermediate values are less common compared to the Gaussian. In other words, natural motion is a *sparse* feature of natural movies.

2.2.2 The timescales of velocity correlations

The optical flow between consecutive frames offers a snapshot of the motion contained in a natural scene; a more complete description should include how this motion evolves over time. Simply analyzing how the optical flow at a given pixel location changes over time is not particularly informative since, by definition, a moving object does not stay in one location for very long. Rather, in order to measure meaningful statistics of object motion we must compute them *along the trajectories of moving objects*. In short, to describe object motion we need to track moving objects.

Object tracking has a long history within the field of computer vision [60]. While many good tracking algorithms exist, their performance depends heavily on the specific application (e.g., single objects which are well-isolated are much easier to track than a clutter of multiple overlapping objects), and finding a general, all-purpose solution is still an area of active research. Fortunately, since we are interested in a statistical description of object motion rather than tracking objects *per se*, we can largely ignore many of the challenges that traditional tracking algorithms face, and instead focus on extending our pixel-level optical

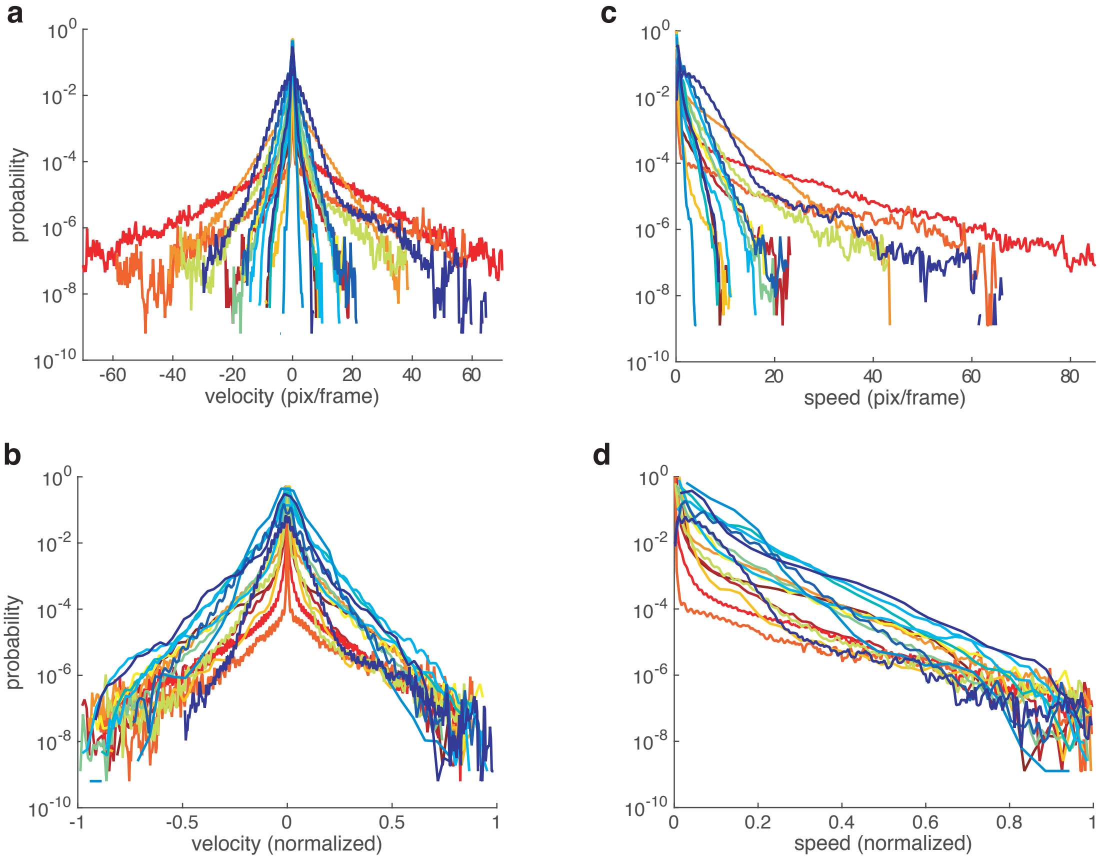

Figure 2.3: **Optical flow histograms.** Colors correspond to movies in Figure 2.1. **a.** Histogram of pooled horizontal and vertical velocity components. **b.** As in **a**, normalized by the peak velocity for each movie. **c.** Histogram of speed (magnitude of velocity). **d.** As in **c**, normalized by the peak speed for each movie.

flow description through time.

To this end we developed a simple *pixel tracking* algorithm that essentially links optical flow vectors across frames. Similar approaches have been taken in [1, 29, 14]. Inspired by fluid dynamics experiments in which easily tracked particles are used to measure complex fluid flows, we instead apply our estimated optical flow fields to an initial grid of fictive "particles" in an iterative fashion (**Figure 2.4**).

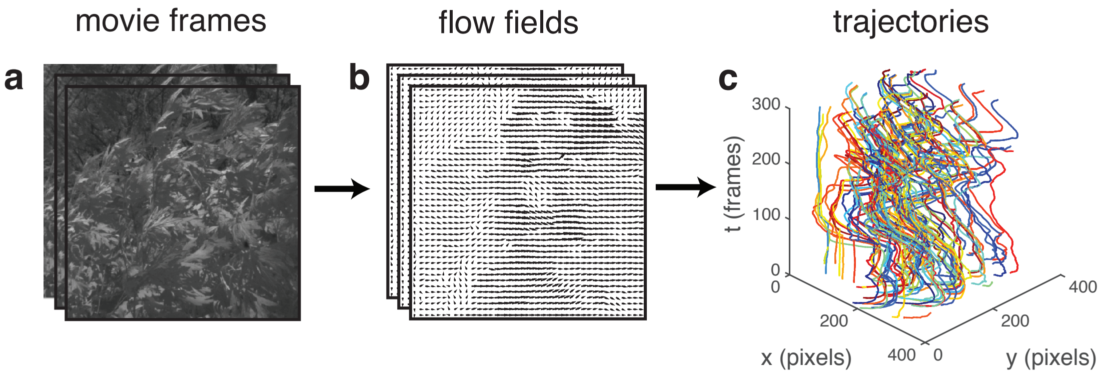

Figure 2.4: **Particle tracking schematic. a.** A stack of movie frames. **b.** A stack of optical flow fields, computed for each pair of consecutive frames. **c.** A set of trajectories obtained by applying flow fields sequentially to a set of "particles" then subsampling, as described in the text.

Once we have some pixel trajectories, how should we analyze them? In the Langevin description of a diffusion process, the velocity of a particle is given by an Ornstein-Uhlenbeck (O-U) process—a stationary Gauss-Markov process characterized by an exponential correlation function (see [22] for a review). This captures the physical scenario of a massive particle, subject to drag, undergoing random velocity kicks. The temporal extent of the correlation function summarizes how predictable the process is. We estimated the velocity, speed, and angle correlation functions of trajectories from each movie clip, shown in **Figure 2.5a-c**. Correlation functions vary dramatically from movie to movie. Some take on negative values, which is inconsistent with a simple O-U process model and suggests an oscillatory component to the motion. In this case, a better model would be the damped harmonic oscillator driven by noise [39], as we describe in Chapter 3. Zooming in on the first several hundred milliseconds, we see significant correlations for at least 100 ms. Since errors in tracking likely lead to an underestimate of correlation, we conclude that natural motion trajectories are highly predictable on a timescale of hundreds of milliseconds.

We also characterized these sets of trajectories as we would a 2-D diffusion process. This kind of analysis has been applied to describe the motion of particles within cells, to infer, for example, whether a particle is undergoing active transport [16, 13] or is being slowed down due to crowding [23]. While the absolute location of such a process is arbitrary—we assume objects positions are equally likely across space (an assumption that may not hold when incorporating behaviors like eye movements)—the displacement over time relative to an initial position provides a compact description of the process. In particular, the *mean squared displacement* grows as a power law function with time, the exponent of which (corresponding to the slope on a log-log plot) measures the diffusivity of the process. An exponent of one is predicted for ideal Brownian motion, in which the velocity is uncorrelated over time. An exponent lower than one indicates a *subdiffusive* process (analogous to crowding), while an exponent greater than one indicates a *superdiffusive* process (analogous to active transport). In the limit of constant velocity, called *ballistic* motion, displacement grows linearly with time, so the exponent will be two. We find mean squared displacement lines with slopes between one and two (**Figure 2.5d**), indicating that natural object motion is in the superdiffusive regime.

2.3 Methods

2.3.1 Movie database collection

We analyzed natural movies from the Chicago Motion Database, which was created for the purposes of statistical analysis and for use as stimuli in neuroscience experiments. Subjects were chosen not only to provide a broad sampling of different moving objects encountered in the natural world, but also to provide sustained motion for relatively long periods of

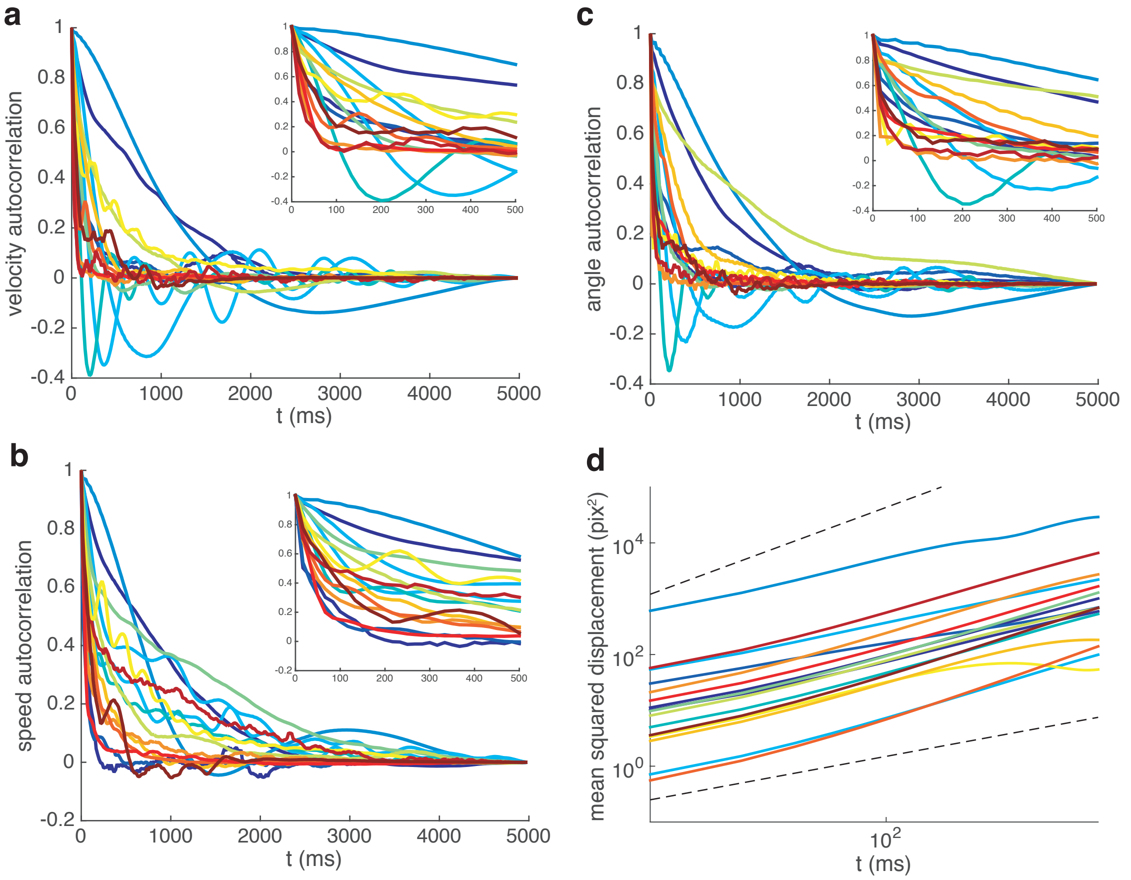

Figure 2.5: **Trajectory statistics.** Colors correspond to movies in Figure 2.1. Insets in **a-c** show the same data with an expanded time axis to emphasize short timescales. **a.** Velocity autocorrelation functions. **b.** Speed autocorrelation functions. **c.** Angle autocorrelation functions. **d.** Mean-squared displacement.

time within a fixed field of view. Subjects the met these criteria tended to be groups of insects and other animals, plants blowing in the wind, and moving water. We chose a subset of movies from the larger database for which our tracking algorithm performed reasonably well. Movies were filmed at either 60 or 120 Hz; we subsampled the 120 Hz movies to 60 Hz to facilitate comparison. Though some movies had three color channels, this information was not used by the optical flow algorithm, which simply converts color inputs to grayscale.

2.3.2 Optical flow algorithm

We calculated the optical flow between pairs of frames using a standard algorithm [52]. The algorithm is a modern implementation of the classic gradient-based optical flow method introduced by Horn and Schunk [25]. Briefly, the aim of the optical flow algorithm is to find a vector field describing the local translation of image intensity that satisfies the *brightness constancy* constraint, the assumption that the luminance at a given point on a moving object does not change over time. Let $E(x, y, t)$ be a continuous, differentiable function of space and time representing the image; x and y are to be interpreted as the position of a point on a moving texture, so that they also depend upon t. Brightness constancy means that the *total derivative* of E with respect to t at a given point should be zero:

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\partial E}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial E}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t} + \frac{\partial E}{\partial t} = 0. \tag{2.1}$$

The time derivatives of horizontal and vertical position introduced above are the optical flow at that position:

$$u = \frac{\mathrm{d}x}{\mathrm{d}t} \text{ and } v = \frac{\mathrm{d}y}{\mathrm{d}t}. \tag{2.2}$$

Since brightness constancy gives us only one constraint with two unknown variables, we must introduce additional constraints to solve for u and v. Optical flow fields are expected to vary smoothly in space except at discontinuities caused by edges, so a natural constraint is to minimize the Laplacian of the optical flow,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \text{ and } \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}. \tag{2.3}$$

In practice, the algorithm minimizes of an objective function consisting of discrete approximations to the brightness constancy and local smoothness constraints (appropriately weighted) and an additional nonlocal smoothness constraint that acts as a median filter. It finds the minimum by jointly varying u and v at each grid location, using a hierarchical pyramid and other techniques from the optical flow literature. See Sun, Roth, and Black (2010) [52] for further details. We used the freely available Matlab code on default settings ('Classic+NL').

2.3.3 Velocity statistics

We computed the speed, s, and direction, θ, by taking the polar transformation of the velocity:

$$s = \sqrt{u^2 + v^2} \text{ and } \theta = \text{atan2}(v, u). \tag{2.4}$$

The function atan2(y, x) returns the angle in radians between the positive x-axis and the coordinates (x, y):

$$\text{atan2}(y, x) = \begin{cases} \arctan\left(\frac{y}{x}\right) & \text{if } x > 0 \\ \arctan\left(\frac{y}{x}\right) + \pi & \text{if } x < 0 \text{ and } y \geq 0 \\ \arctan\left(\frac{y}{x}\right) - \pi & \text{if } x < 0 \text{ and } y < 0 \\ +\frac{\pi}{2} & \text{if } x = 0 \text{ and } y > 0 \\ -\frac{\pi}{2} & \text{if } x = 0 \text{ and } y < 0 \\ \text{undefined} & \text{if } x = 0 \text{ and } y = 0. \end{cases} \tag{2.5}$$

Velocity and speed histograms were calculated using 0.25 pix/frame bin sizes and normalized by the total number of samples.

2.3.4 Pixel tracking algorithm

Starting from a grid of initial horizontal and vertical positions $(\mathbf{x}_0, \mathbf{y}_0)$ (bold face indicates vectors where each element corresponds to a different trajectory), the position of each particle is updated recursively according to

$$(\mathbf{x}_t, \mathbf{y}_t) = (\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + (\mathbf{u}_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), \mathbf{v}_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})), \tag{2.6}$$

where $(\mathbf{u}_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), \mathbf{v}_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}))$ is the optical flow between frames $t-1$ and t (with frame indices starting at zero) evaluated at the previous positions $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$. Since the optical flow takes continuous values, so do the positions, and we must estimate the optical flow off of grid locations using bilinear interpolation. Trajectories end when particles exit the field of view. This method also tracks the stationary background, which will only decrease the statistics we measure. We remove stationary trajectories by detecting runs of 5 or more frames with no motion (speed less than 0.1 pix/frame) and eliminating them (keeping only the first segment of the trajectory if it is interrupted). We are left with a dense covering of the moving objects in the scene. These trajectories are highly spatially correlated and therefore redundant, so we subsample them, weighting each by the total distance traveled to encourage long, continuous trajectories. Since errors in this integration process accumulate over time, we limit ourselves to time windows of 300 frames (note that trajectories may end sooner), and repeat the process in increments of 150 frames, covering the length of each clip with 50% overlap to ensure that all moving objects are well-sampled.

In addition to the inevitable accumulation of measurement errors from the optical flow, more catastrophic errors can occur when particles "fall off" of or are "picked up" by moving objects, so that a trajectory no longer corresponds to the motion of a single object. The algorithm works best for moderate speeds and relatively large objects; objects that move too quickly or are too small tend not to be linked properly across successive frames. There is also conceivably an "aperture effect" due to the fact that faster objects simply do not remain in the frame as long as slower objects. For these reasons we suspect that this algorithm generally leads to underestimation of the statistics we calculate, especially at longer time delays, which should be considered as a lower bound of the true values.

2.3.5 Trajectory statistics

The mean squared displacement is simply the average Euclidean distance between some initial position at time t and the position along the same trajectory τ seconds later:

$$\mathrm{MSD}(\tau) = \langle \sqrt{[x(t) - x(t+\tau)]^2 + [y(t) - y(t+\tau)]^2} \rangle, \tag{2.7}$$

where the average is taken over trajectories and time.

The autocorrelation function of a time-varying (linear) quantity $q(t)$ with mean μ_q is

$$C_q(\tau) = \langle [q(t) - \mu_q][q(t+\tau) - \mu_q] \rangle. \tag{2.8}$$

We use this formula to calculate the velocity (pooling horizontal and vertical components) and speed autocorrelation functions.

For the angle θ, which is a circular quantity, we use the formula

$$C_\theta(\tau) = \langle \cos[\theta(t) - \theta(t+\tau)] \rangle. \tag{2.9}$$

2.4 Discussion

Optical flow distributions have previously been measured in the context of camera motion applied to range images in [46, 15]; like object motion, this self-generated motion also follows a kurtotic distribution. Kurtotic distributions are a surprisingly common feature of natural image statistics. For example, Gabor filters, a classic model of the edge-detector-like receptive field properties of V1 simple cells, produce a kurtotic distribution of responses when applied to natural images; this fact underlies the influential theory of sparse coding [21, 40]. Intuitively, motion corresponds to a spatiotemporal edge, and spatiotemporal edge-detector filters, such as those derived from independent component analysis of natural movies [55], also yield kurtotic response distributions.

The chief contribution of our work, showing that object velocity tends to be highly correlated for (conservatively) hundreds of milliseconds, is potentially relevant to a number of research areas within neuroscience. Our primary aim was to inform the design of object localization experiments in the retina [33, 42] such as those described in the next chapter. Beyond sensory encoding, realistic object trajectories with natural statistics could be useful in sensorimotor tasks, such as prey capture [11, 35], object tracking, or reach and grasp tasks. By extension of the efficient coding argument, we expect sensorimotor systems should be tuned to natural trajectories as well.

Our sample of movies is by no means exhaustive, and we do not attempt to find average statistics describing the entire ensemble of natural movies. Rather, we highlight the diversity of autocorrelation functions found across subjects. Depending on an animal's behavioral and ecological niche, it may be tuned to different motion statistics, which future work could help to disentangle. We emphasize that natural motion seems to be correlated for *minimum* of 100 ms, which should be broadly applicable to neural systems.

CHAPTER 3

OPTIMAL PREDICTION OF A MOVING BAR

3.1 Introduction

Ultimately we would like to test whether the retina is optimized for prediction of the natural stimuli it evolved to encode. However, natural movies are complex and potentially contain many different moving features, in addition to a rich and complicated spatial structure. Here we simplify matters considerably by using artificial stimuli consisting of a single feature—the position of a moving bar. Our precise control of the stimulus allows us to calculate the information neural responses contain about past and future trajectories. We can then compare these values to the theoretical optimum using the information bottleneck method to directly test whether the retina is optimized for prediction.

In [42], the authors demonstrated optimality for a particular set of stimulus statistics. We follow up on this result by presenting two sets of stimulus statistics to the same retina for direct comparison. We find qualitatively different results for the two sets of stimuli, attesting to the importance of the natural motion statistics described in the previous in shaping the predictive capabilities of the retina.

3.2 Results

Not all features of visual input are predictable, and optimal prediction requires selectively encoding predictable features to make the most efficient use of limited resources. For this reason, we use a carefully designed stimulus that contains both predictable and unpredictable components. The stimulus consisted of a white bar 100 µm wide and stretching the length of the array on a black background (**Figure 3.1a**). The position x_t and velocity v_t were updated according to the following stochastic difference equations:

$$v_{t+\Delta\tau} = [1 - \Gamma\Delta\tau]\, v_t - \omega^2 x_t \Delta\tau + \xi_t \sqrt{D\Delta\tau}, \tag{3.1}$$

$$x_{t+\Delta\tau} = x_t + v_t \Delta\tau, \tag{3.2}$$

where $\Delta\tau = 1/120$ s is the update time of the display, $D = 2.7 \times 10^6$ pixel^2/s^3 is a fixed diffusion constant, ξ_t is a Gaussian white noise process with unit variance, and Γ and ω are parameters controlling the drag and spring force. Taken together, the equations simulate the one dimensional position of a massive particle undergoing diffusive motion, but tethered to the center of the display. By fixing $\Gamma_1 = 20$ s^{-1}, $\Gamma_2 = 4$ s^{-1} and $\omega_1 = 1.5 \times 2\pi$ rad s^{-1}, $\omega_2 = 0.67 \times 2\pi$ rad s^{-1}, we define two statistical environments, characterized by their position and velocity autocorrelation functions (**Figure 3.1b,c**). Intuitively, the time it takes the autocorrelation functions to decay to zero determines how far into the future one could predict the stimulus given perfect knowledge of its current position and velocity. Examples of bar position trajectories for the two statistical environments are shown in **Figure 3.1e,f**, which also illustrate the trial structure of the experiment: In each trial a unique past trajectory converges onto one of a collection of common future trajectories. This allows us to calculate the information the neural responses contain about the future before it happens, based on responses to past trajectories that are all unique yet statistically linked to the common future trajectory. We presented 100 unique pasts for each of 30 common futures for a total of 3,000 trials for each statistical environment. For visualization purposes we show only 2 common futures with 10 unique pasts each. The midpoint of each trial, designated $t = 0$, marks the point of convergence of past and future.

3.2.1 Stimulus dependence of bound saturation

We calculate the amount of information small groups of cells, randomly sampled from the population, carry about the identity of the common future trajectory as a function of time

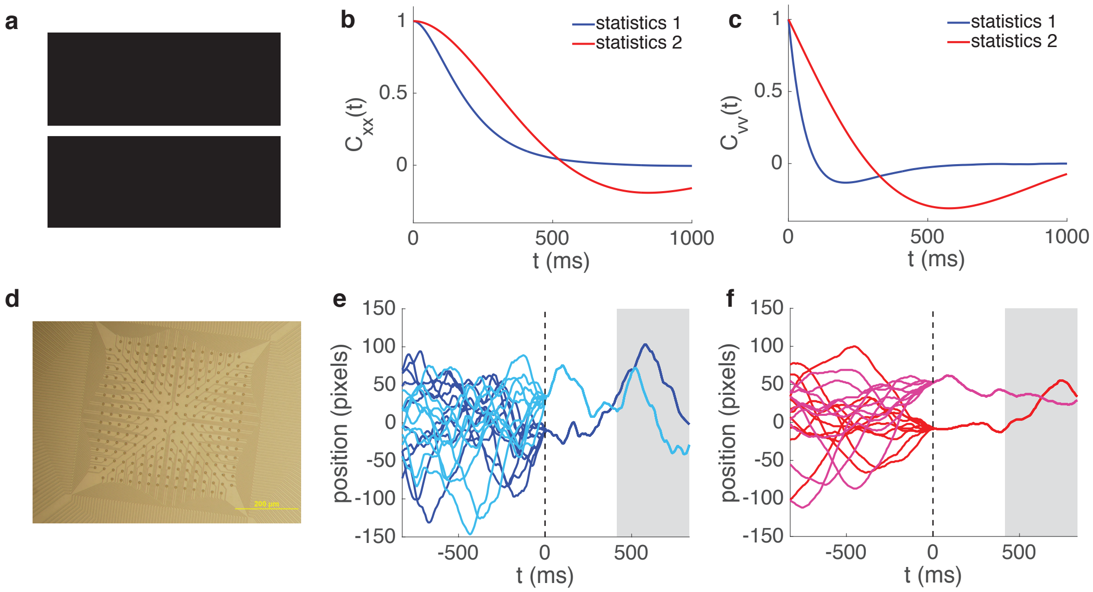

Figure 3.1: **Stimulus design. a.** Example stimulus frame. **b.** Position autocorrelation functions. **c.** Velocity autocorrelation functions. **d.** Image of a 252-channel multi-electrode array, courtesy of Olivier Marre. **e,f.** Example stimulus position traces for statistics 1 (**e**) and 2 (**f**), illustrating the common future experimental design: 20 unique past trajectories converge onto 2 common future trajectories (10 each). The experiment contained 30 common futures each with 100 unique pasts for each statistical environment; only a subset are shown for display purposes.

relative to the convergence time $t = 0$. We take positive Δt to represent time before convergence. Hence, large negative values correspond to time points long after convergence, and we average the last 25 time points to calculate the "information about the past" (I^{past}). For "information about the future" (I^{future}) we use the information calculated using responses just before the convergence, at $\Delta t = 8.33$ ms. In **Figure 3.2a,b** we plot these information values for a single 4-cell group. For statistics 1, we see a steep decline in information with delay, while for statistics 2 the decline is more gradual. This trend is confirmed when we calculate the average information for 1000 such randomly sampled 4-cell groups (**Figure 3.2c,d**). When we compare information values directly across statistical environments, the results are clear: We almost always find more information about the past for statistics 1 (**Figure 3.2e**), yet more information about the future for statistics 2 (**Figure 3.2f**).

To place these information values in context we compare them to the theoretical bound on information about the future (a function of information about the past) that is set by the stimulus statistics. In the region of small values of I^{past} this bound is approximately the unity line, reflecting the simple fact that the response cannot contain more information about the future than the past. We find that information values for all group sizes cluster just below the bound for the second statistical environment, but fall significantly short for first (**Figure 3.3a,b**). We can also find individual groups of cells (an example group is denoted by a red circle in **Figure 3.3a,b**) that saturate the bound calculated for a fixed value of I^{past} with a delay Δt, again for the second statistical environment but not the first (**Figure 3.3c,d**). This is a very robust indication that the retina optimally encodes the predictive features of the stimulus for statistics 2, allowing it to predict the stimulus into the future as far as the stimulus statistics will allow. On the other hand, for statistics 1, the retina is still informative about the past stimulus (in fact, moreso than for statistics 2), but this information is not useful for predicting the future.

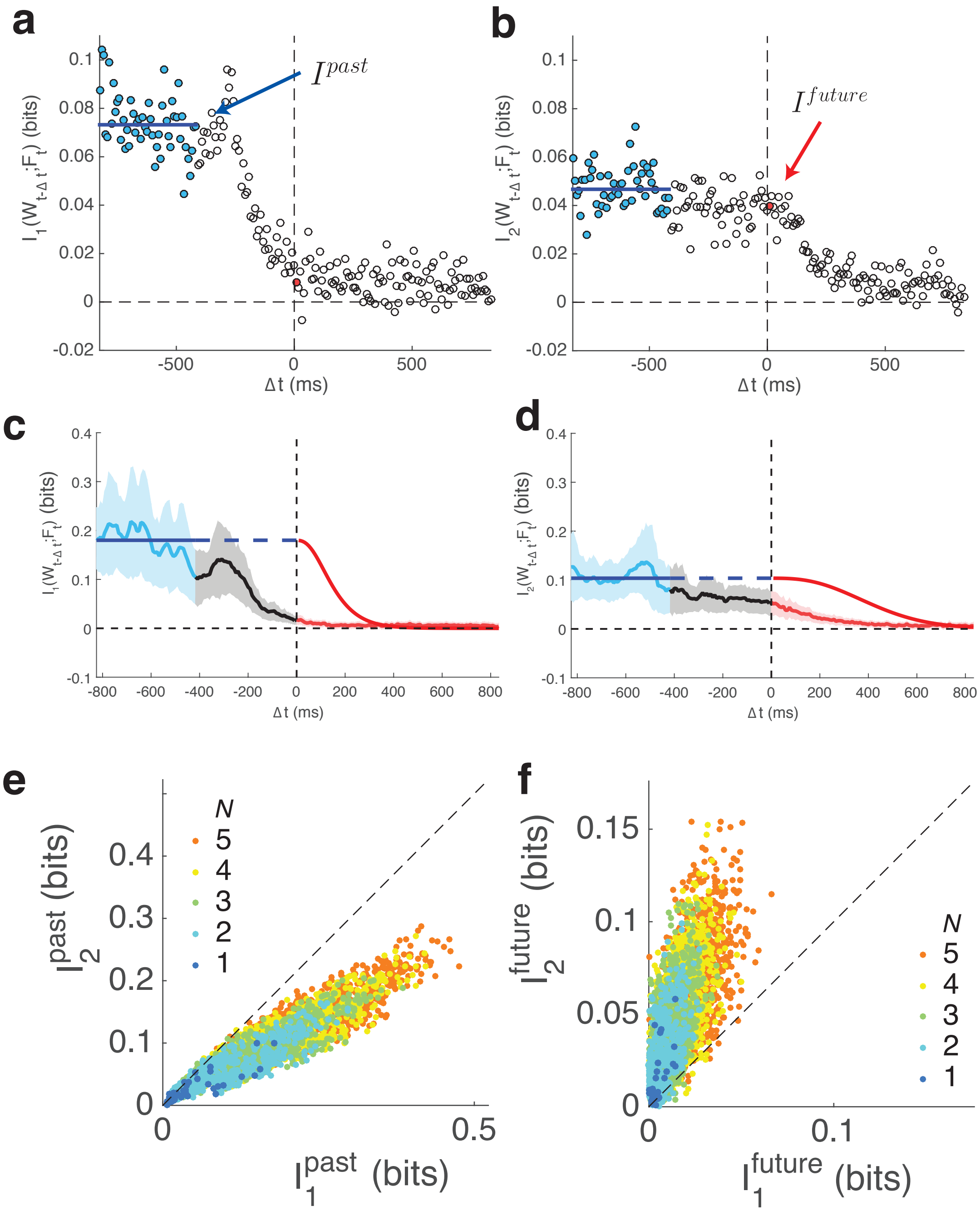

Figure 3.2: **Past and future information. a,b.** Information as a function of delay for a single 4-cell group, for statistics 1 (**a**) and 2 (**b**). I^{past} is defined as the average value (blue line) of information at the last 25 time points (blue dots), while I^{future} is defined as the information at the time point immediately before the trajectories converge (red dot). **c,d.** Mean and standard deviation of information over 1000 4-cell groups for statistics 1 (**c**) and 2 (**d**). The information bound (red line) corresponding to the average information about the past (blue line) is shown. **e,f.** Comparison of I^{past} (**e**) and I^{future} (**f**) across statistics for groups of different sizes.

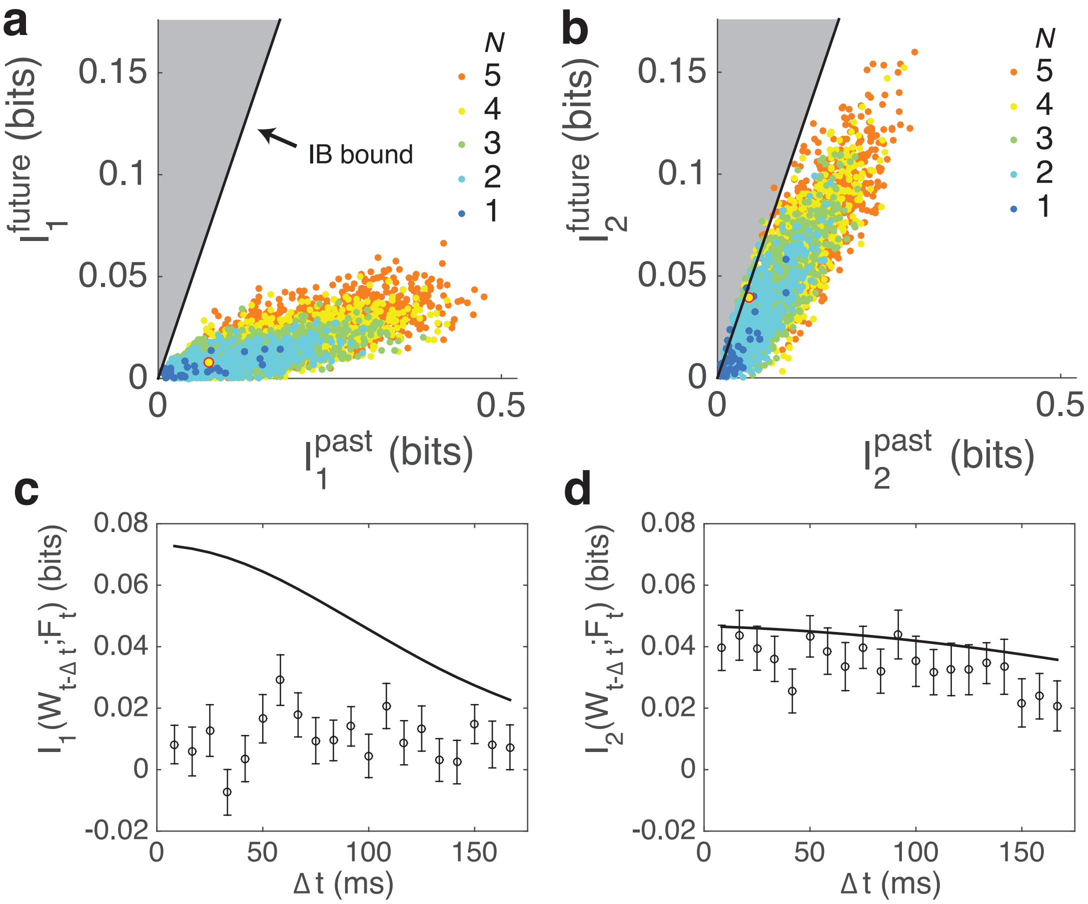

Figure 3.3: **Bound saturation. a,b.** Comparison of I^{past} and I^{future} for statistics 1 (**a**) and 2 (**b**). The bound (black line) derived from the information bottleneck problem is just the unity line for small values of information. **c,d.** Information as a function of delay for a single 4-cell group (circled in red in **a,b**), for statistics 1 (**a**) and 2 (**b**), with the bound also computed as a function of delay (black line) for comparison.

3.2.2 Shifts in effective delay

How is it that the retina is capable of optimal prediction in one statistical environment and not the other? To gain some insight, we calculated the amount of information the neural response contains about the instantaneous bar position as function of delay, rather than entire past and future trajectories. Across group sizes, we find a robust shift in the peaks of these curves (which we interpret as the effective delay of the system) as we change statistical environments (**Figure 3.4a,b**). In particular, the distribution of effective delays for statistics two is broader and its median shifted forward from -100 ms to -66.7 ms (**Figure 3.4c**).

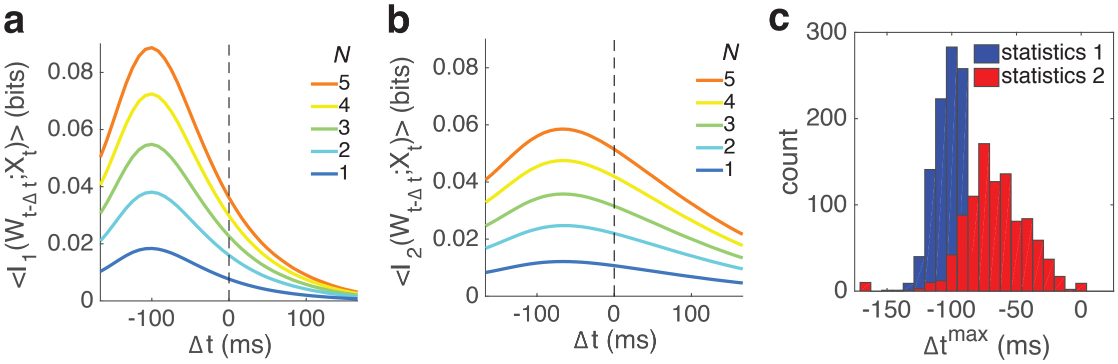

Figure 3.4: **Instantaneous information and effective delay. a,b.** Instantaneous position information as a function of delay for different group sizes for statistics 1 (**a**) and 2 (**b**). **c.** Histogram of effective delays calculated for 1000 4-cell groups.

3.2.3 Biphasic filter model

What can account for the robust shift in effective delay with stimulus statistics? It seems reasonable that mechanisms underlying motion anticipation are engaged more in the second statistical environment, but which elements are affected? Are adaptation effects required? To answer this, we started with the most basic model of ganglion cell responses, the linear filter. Temporal linear filters, calculated either as a function of contrast from a white noise checkerboard stimulus or as a function of stimulus position using ridge regression, tended to have a biphasic shape, illustrated for an example neuron in **Figure 3.5a**. Note that for a space-time separable linear filter, the transformation from position to filter output consists of a spatial activation (the overlap of the bar with the spatial receptive field) convolved with the temporal filter, so the overall transformation from bar position to neural response inherits this biphasic shape. This example neuron also shows a shift in delay for different statistics, in this case achieving zero effective delay for statistics 2 (**Figure 3.5b**). We can easily achieve a similar shift in delay with a model biphasic filter (**Figure 3.5d,e**). The intuition for constructing such a filter, and why it produces a shift in delay, is as follows: We can conceptualize a filter consisting of a Gaussian bump at some delay as reporting the position at that delay, with some temporal smoothing to eliminate noise (**Figure 3.5d**, red curve). This results in an information curve with peak at that delay (**Figure 3.5c**, red curve). Similarly, a sum of Gaussians with opposite signs gives us the change in the two delayed positions, which is proportional to the velocity (**Figure 3.5d**, green curve). This tells us nothing about the position at the midpoint of the two delays, but it is informative about future and past positions because, e.g., a positive velocity means that the future position is more likely to be positive and the past negative (**Figure 3.5c**, green curve). Taking a linear combination of these "position" and "velocity" filters with an appropriate scaling (**Figure 3.5d**, black curve), which can be thought of intuitively as finding the position and adding the velocity multiplied by the time step, results in a filter with zero effective delay for statistics 2 and a nonzero delay for statistics 1 (**Figure 3.5e**). Thus, the basic phenomenology of effective delay changing with stimulus statistics can be accounted for by the biphasic nature of the neural encoding, although we do not rule out the possibility of adaptation or other nonlinear effects playing a role.

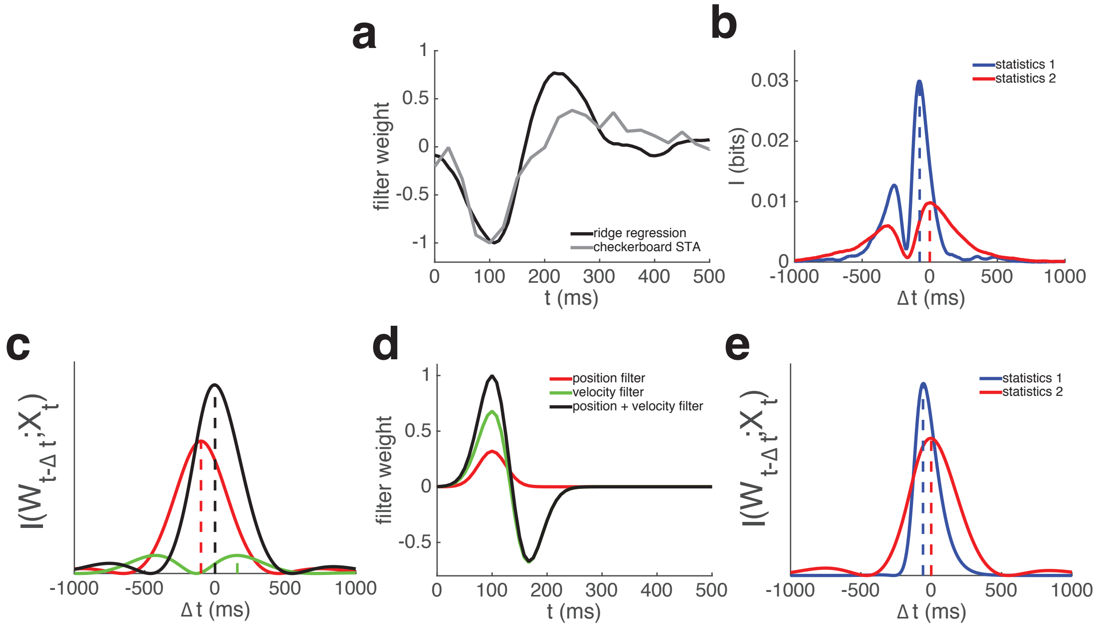

Figure 3.5: **Prediction with biphasic linear filters. a.** Linear filters calculated using ridge regression for the moving bar stimulus (a function of position) and by simple spike-triggered averaging of a checkerboard stimulus (a function of contrast) for an example neuron. **b.** Instantaneous position information for the neuron in a for statistics 1 and 2. **c.** Instantaneous information for the three filters in **d** for statistics 2. **d.** Model filters for estimating position (a Gaussian bump), velocity (a difference of Gaussians with equal positive and negative weight), and future position ("position + velocity") (a difference of Gaussians with more positive weight than negative). **e.** Instantaneous position information for the "position + velocity" filter for statistics 1 and 2, demonstrating the decrease in effective delay seen in **b**.

3.3 Methods

3.3.1 Experimental protocol

Experiments were carried out at Institut de la Vision in Paris, France, under the supervision of Olivier Marre, and in accordance with institutional animal care and use requirements. Adult Long-Evans rats were rendered unconscious using a CO_2 chamber, then killed via cervical dislocation. The eyes were removed and transferred to a bath of Ames' solution. In a dark room using low red light illumination, eyes were hemisected and cleared of vitreous to expose the ganglion cell layer of the retina. A piece of retina slightly larger than the recording array was cut using a scalpel and transferred from the sclera to a perforated dialysis membrane stretched over a small ring, with the ganglion cell layer up. This ring was then attached to a shaft and carefully lowered onto a multi-electrode array (MEA) for recording. The MEA is embedded in glass with cylindrical dish on top, allowing the retina to be perfused with oxygenated Ames' solution during recording.

3.3.2 Recording and spike sorting

The MEA (Multi Channel Systems) consists of 252 electrodes, arranged in a 16 by 16 grid with 4 corners missing, with 60 µm spacing between electrodes. The raw voltage was digitized and saved at a rate of 20 kHz. Spikes from individual neurons were inferred from the data using the spike sorting algorithm described in [32]. Due to the short distance between electrodes, action potentials from individual retinal ganglion cells cause voltage changes across multiple electrode; if two cells spike simultaneously, the resulting waveform will be an approximate sum of the waveforms produced by individual spikes. Spiking events are detected by simple thresholding. The algorithm first identifies well-isolated spikes and clusters their waveforms (across all electrodes) into a set of templates representing putative neurons. The templates are then fit to overlapping spikes in a greedy fashion: the best fitting template is assigned first and is subtracted from the data, then another template is fit to the residuals and subtracted, and the process is iterated until no spikes remain.

Spikes from putative neurons were post-processed by hand to remove spurious low-amplitude spikes fit by the algorithm. Cells were rejected if they had a high number of refractory period violations (interspike intervals less than 5 ms) or were unstable across the recording. A total of 48 neurons were included in the data set.

3.3.3 Stimulus design

The stimulus was presented using a digital micromirror device (DMD), illuminated by a white light source with an ND3 filter, and projected onto the photoreceptor layer of the retina. We used a minimal amount of light in order to avoid bleaching of photopigments and achieve longer recording times. Each pixel of the DMD corresponded to $\sim 3.3\,\mu\text{m}$ on the retina. Stimulus frames were updated at a rate of 120 Hz.

The stimulus consisted of a white bar, 29 pixels wide and extending the length of the recording area, on a black background. It was animated according to the equation above, which corresponds to a physical model of a damped harmonic oscillator driven by noise [39].

3.3.4 Mutual information

We used the Bayesian entropy estimator described in [3] to calculate the mutual information between stimulus and response. The estimator is similar to the NSB estimator [36] but uses a prior distribution that is more appropriate for sparse neural data. We used the formula $I(X;Y) = H(X) + H(Y) - H(X,Y)$, rather than calculating the conditional entropy, since it shows better convergence properties in simulations [2]. Here, the response consisted of binarized neural "words"—patterns of spikes and silences for a group of neurons in a given 8.33 ms time bin (using stimulus frame times as bin edges). The stimulus was either the *identity* of the common future for a trial, to calculate I^{past} and I^{future}, or the position of the bar, quantized to 32 approximately uniformly distributed values, to calculate the instantaneous information. 1000 groups of N cells, where N is one to five, were randomly subsampled from the population. We limited ourselves to five cell groups, since larger groups lead to nonzero information values after shuffling to break the association between stimulus and response—an indication of bias due to limited data.

3.3.5 Information bottleneck

The information bottleneck method [53] allows us to calculate the maximum amount of information about the future that neural responses can contain for a given amount of information about the past. Since the stimulus positions are jointly Gaussian, we can use the analytic solution derived in [17]. The solution is just a function of the covariance matrix of past and future positions, which we calculate from the simulated trajectory.

3.4 Discussion

These results provide strong evidence for the hypothesis that the retina is optimized for prediction, reinforcing the results of [42], but with the caveat that optimal prediction depends crucially on the statistics of motion governing the stimulus. While [42] demonstrated saturation of the bound for some groups of cells, the results presented here for statistics 2 are more robust, with the bulk of randomly sampled groups sitting close. Notably, the stimulus used in [42] followed statistics 1, here. This is suggestive of a species difference between salamanders and rats, in which salamanders are tuned to motion correlations on shorter time scales than rats, reflecting their different environments and behaviors. However, the difference may also be attributed to subtle differences in the stimulus, namely the overall light level and the polarity of the bar ([42] used a black bar on a gray background rather than the white bar on black background used here). Further experiments will be required to determine whether this is in fact a species difference, as well as to determine precisely what statistics lead to optimal prediction.

In the past, motion anticipation in the retina (the deterministic equivalent to the statistical results shown here), has been attributed, phenomenologically, to a contrast gain control mechanism [8, 30]. More biologically realistic models implicate shunting inhibition [27] and, for a particular RGC subtype capable of accommodating a wide range of velocities, gap junctions [54]. Regardless of the details of the mechanism, we emphasize that, from a statistical point of view, the essential features of prediction are captured by a biphasic linear filter. Above we give an intuitive explanation for how biphasic filters can lead to prediction; a more rigorous treatment of this phenomenon involves calculating the group delay (the derivative of the phase with respect to frequency) of the filter [58]. Surprisingly simple circuit elements can give rise to a negative group delay and hence prediction; in [57] the author uses a simple leaky integrator with delayed feedback inhibition, but we suspect that the largely feedforward elements of the retina, with different delays, could achieve similar results. This will be the subject of future work.

One limitation of our results is that, for the fairly small information values we can calculate directly for small groups of cells, the bound given by the information bottleneck for the stimulus is just the unity line. For larger values of past information, the slope gradually decreases, until it plateaus at some maximum value. More convincing evidence of bound saturation would come from observing values near the "knee" of this curve (values at the plateau would indicate that the retina is investing too many bits of information about the past without gaining any additional information about the future). Future work could address this issue by approximating the information for larger groups of cells, using, for example, linear decoding [59, 33].

CHAPTER 4

INTERNAL PREDICTIVE INFORMATION

4.1 Introduction

The difference between efficient coding and optimal prediction boils down to specifying an optimization problem for the retina to solve. Efficient coding is formulated simply as maximizing stimulus information given limited neural resources (the total entropy of the response). Optimal prediction can be formulated as an information bottleneck (IB) problem, which seeks to compress the stimulus as much as possible while retaining information about a second relevance variable, in this case the future stimulus. In both cases, the solution depends critically on the statistical structure of the stimulus.

How do we differentiate between these two coding schemes, particularly for the complex natural stimuli the brain has evolved to process? We approach this problem by examining both the statistical structure of a set of natural movies, with an emphasis on their motion content, and the responses of a population of retinal ganglion cells being stimulated with these movies. In doing so, we hope to shed light on the statistical dependencies between past and future stimulus and response.

4.2 Theory

4.2.1 Experimental predictions of competing theories

Given the complexity of both the structure of natural movies and their transformation into neural spike trains, how can we hope to determine whether the retina is optimized for prediction or merely efficiency? One approach is to analyze the correlation structure of the RGC responses and compare it to what would be expected from an idealized model. The details of this analysis are given in Appendix A. We use the simplest possible model, in which the response R to a Gaussian stimulus X is the convolution of X with a linear filter

A with additive Gaussian noise N. In the frequency domain this is simply:

$$R = AX + N. \tag{4.1}$$

The filter A is then optimized for a particular objective function. For efficient coding, we'd like to minimize the overall magnitude of the response (its variance) while maximizing information transmission. This gives us the objective function

$$\min_{|A|^2} \mathcal{L} = \mathrm{Var}(R) - \beta I(R; X) \tag{4.2}$$

where β is a parameter that determines how much information is transmitted. Because of our simplifying assumptions, this problem has an analytic solution:

$$|A|^2 = \frac{\beta - S_N}{S_X}, \tag{4.3}$$

where S_X and S_N are the power spectra of the input and noise, respectively. This solution is known as a *whitening filter* because it leads to a flat response power spectrum:

$$S_R = \beta. \tag{4.4}$$

For the information bottleneck problem, we instead wish to minimize the information about the stimulus while maximizing information about a *relevance* variable Y:

$$\min_{|A|^2} \mathcal{L} = I(X; R) - \beta I(R; Y). \tag{4.5}$$

This leads to the solution

$$|A|^2 = \frac{S_N}{S_X} \frac{\beta \gamma_{XY}^2 - 1}{1 - \gamma_{XY}^2}. \tag{4.6}$$

Here, A is similar to a whitening filter in that the first term inverts the stimulus power spectrum. However, the second term scales frequencies according the the coherence between X and Y, γ^2_{XY}, leaving an output spectrum that is not flat:

$$S_R = S_N \left(\frac{\beta\gamma^2_{XY} - 1}{1 - \gamma^2_{XY}} + 1 \right). \tag{4.7}$$

In the case of prediction of natural movies, Y is the future stimulus, and we expect γ^2_{XY} to be large at low frequencies, and hence the the responses should be significantly correlated.

Technically, the analysis above is incomplete for the past-future information bottleneck problem because it does not require that the solution be causal. A causal solution is one in which the filter is only a function of past input. This requirement makes the analysis much less straightforward (although it can still be solved numerically). Since our goal is simply to provide intuition for what to expect in the correlation structure of neural responses for the two theories, we end here with the conclusion that efficient coding leads to decorrelated responses, while for optimal prediction the responses should have correlations that reflect the predictable features of the stimulus.

4.2.2 Autoinformation

In what follows we make use of the *autoinformation*, the mutual information between neural responses in two time bins separated by a delay τ. To better understand this quantity, we derive an analytical approximation for it that is valid for sufficiently small bin sizes under suitable assumptions. The details are given in Appendix B. We assume the spikes of each neuron i are generated by an inhomogeneous Poisson process with rate $r_i(t)$, with mean rate $\bar{r}_i$ and auto- and cross-correlation functions $\rho_{ij}(\tau)$. The mutual information I between binned population spike counts $\mathbf{n}(t)$ decreases with the square of the bin size Δt. We define the limit of the ratio of mutual information to squared bin size:

$$\mathcal{I}(\tau) = \lim_{\Delta t \to 0} \frac{I(\mathbf{N}(t); \mathbf{N}(t+\tau))}{\Delta t^2}. \tag{4.8}$$

For sufficiently small but finite Δt, this allows us to recover the autoinformation by multiplying $\mathcal{I}$ by Δt^2. We take an approach similar to [43] to eliminate terms involving more than two spikes, and hence any higher-order correlation structure. This yields the surprisingly simple result,

$$\mathcal{I}(\tau) = \sum_{i,j} \mathcal{I}_{ij}(\tau), \tag{4.9}$$

where

$$\mathcal{I}_{ij}(\tau) = \rho_{ij}(\tau) \log \frac{\rho_{ij}(\tau)}{\bar{r}_i \bar{r}_j} - (\rho_{ij}(\tau) - \bar{r}_i \bar{r}_j). \tag{4.10}$$

Rearranging, we have

$$\mathcal{I}(\tau) = \sum_i [\mathcal{I}_{ii}(\tau) + \sum_{j \neq i} \mathcal{I}_{ij}(\tau)]. \tag{4.11}$$

This tells us that, for small bin sizes, the autoinformation for a single neuron is a simple function of its firing rate autocorrelation; for groups of neurons, the individual autoinformations sum, along with pairwise contributions that are a function of the cross-correlation. Thus, it is always easier to predict the activity of groups of neurons than it is to predict each neuron's activity individually; the predictability individual neurons add together, and correlations between cells can only add further predictability.

Applying this approximation to neural data will be the subject of future work. Preliminary results suggest that it fits the autoinformation of individual neurons quite well at $\Delta t = 16.7\,\mathrm{ms}$, but overestimates the autoinformation of small groups, suggesting that higher-order correlations still play a large role at this timescale.

4.3 Results

4.3.1 A conserved autoinformation timescale across stimuli

Now that we have gained some intuition about the predictions of competing theories and the autoinformation quantity we will use to assess them, it is time to apply these ideas to neural data. We make use of a data set consisting of 93 retinal ganglion cells from the larval tiger salamander responding to 5 different natural movie stimuli (**Figure 4.1**). In order to learn about the temporal structure of these movies, we apply two different measures: the contrast autocorrelation and the novel velocity autocorrelation described in Chapter 2. We find that contrast correlations extend well beyond 500 ms (**Figure 4.2a**), while velocity correlations are significant for approximately 200 to 500 ms (**Figure 4.2b**).

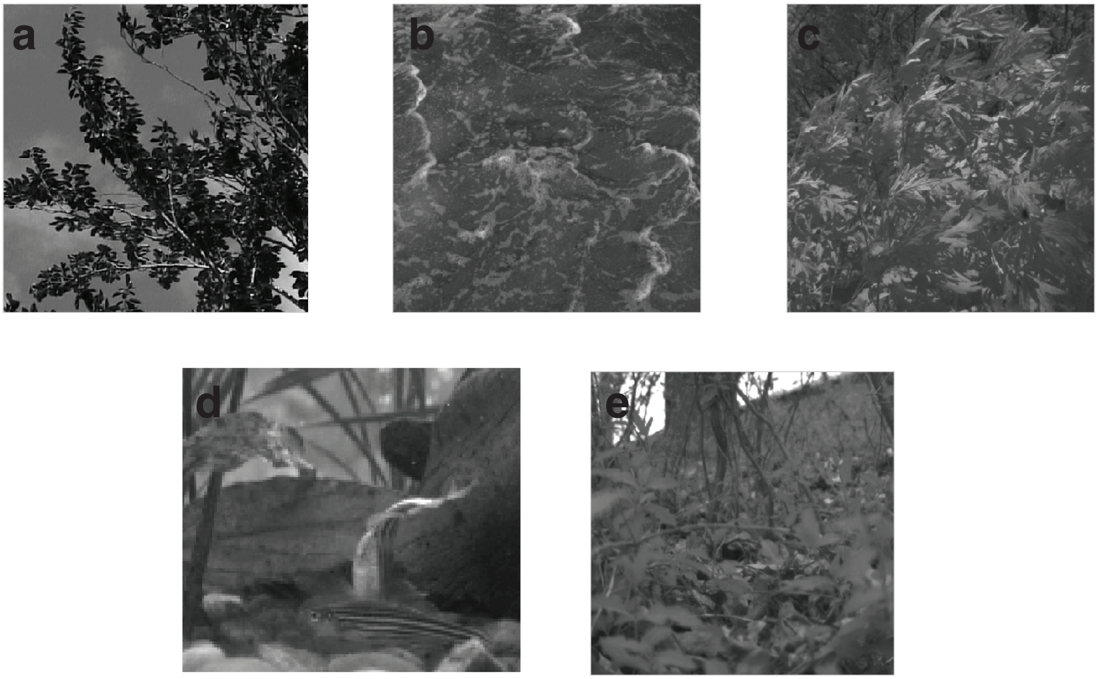

Figure 4.1: **Stimulus frames. a-e.** Example frames from the five natural movie clips used as stimuli. The movies depict (**a**) branches, (**b**) water, (**c**) leaves, (**d**) fish, and (**e**) camera motion through a wooded environment.

We calculate the autoinformation for 1000 5 cell groups, randomly sampled from the population, in response to each of the 5 movies. The average autoinformation is different for each movie (**Figure 4.3a**), but this is only due to the fact that different movies drive responses to different degrees; when we normalize by the overall response entropy for each movie, the curves collapse (**Figure 4.3b**), suggesting that a common level of predictability in the responses is maintained despite drastic changes in the input. We rule out the possibility

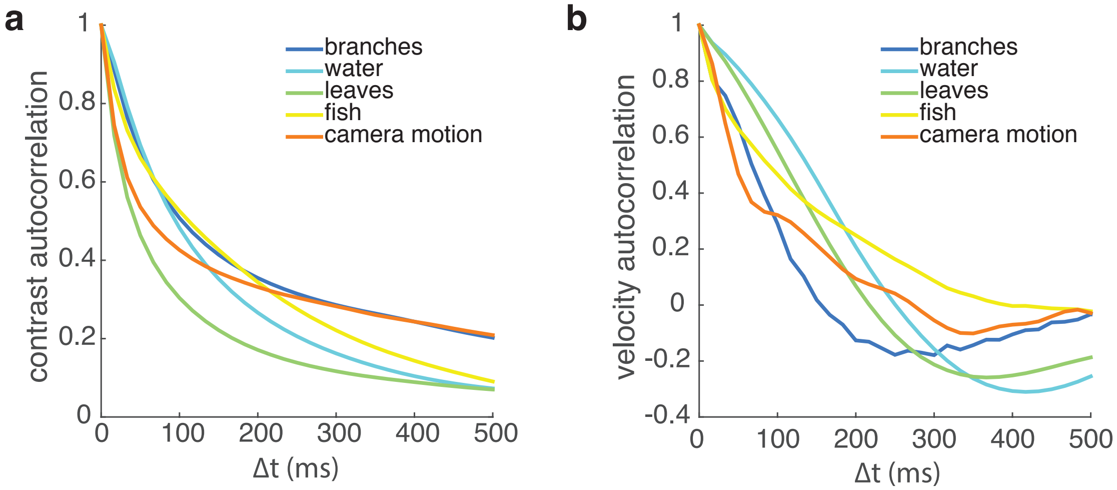

Figure 4.2: **Statistics of natural movie stimuli. a.** Contrast autocorrelation functions. **b.** Velocity autocorrelation functions.

that this predictability is due to the internal dynamics of the retina by calculating the autoinformation in response to a random checkerboard stimulus, which is uncorrelated across frames; we find that in this case the autocorrelation falls off quickly after 33.3 ms, the frame duration of this stimulus. Thus, the response autoinformation seems to reflect the structure of the natural movie input; the timescale of about 200 ms is much shorter than that of the measured contrast autocorrelation, but roughly agrees with the timescale of velocity autocorrelation.

Further evidence that the response autoinformation corresponds to the structure of the input comes from the observation of strong correlations between autoinformation and stimulus information; that is, groups of cells with higher autoinformation also tend to carry more information about the stimulus (**Figure 4.4a**). Is this information actually useful for predicting the future stimulus? While we cannot measure information about the future for natural movies, we can for the common future stimulus paradigm described in Chapter 3. We find that the autoinformation is also highly correlated with information about the future in this data set (**Figure 4.4b**). This is an important consideration given that neurons downstream of retina need to make predictions based entirely on the structure of their input,

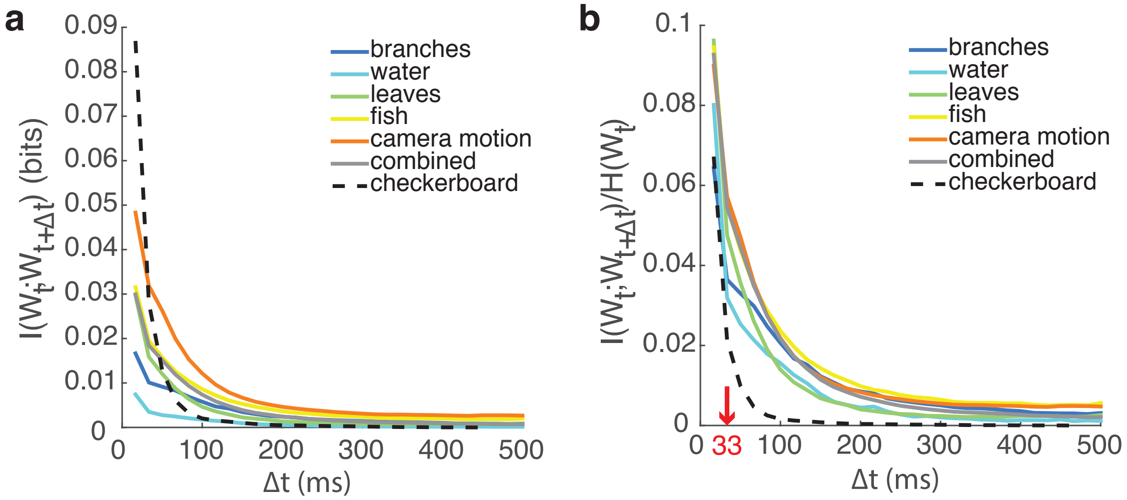

Figure 4.3: **Autoinformation functions. a.** Average autoinformation of 1000 5 cell groups in response to 5 different natural movies and a random checkerboard stimulus. **b.** As in **a**, normalized by entropy.

since they cannot observe the stimulus directly [42].

4.3.2 Competing notions of efficiency

In addition to making different predictions about the correlation structure of responses, efficient coding and optimal prediction also lead different notions of efficiency by which we can measure the performance of the retina. The classical definition of efficiency is simply the fraction of the total response entropy that carries information about the stimulus:

$$E_C = \frac{I(R;X)}{H(X)}. \tag{4.12}$$

On the other hand, the information bottleneck suggests that, given an appropriately defined relevance variable Y, efficiency should measure how close the relevant information $I(R;Y)$ comes to the theoretical maximum $I(R^*;Y)$,

$$E_{IB} = \frac{I(R;Y)}{I(R^*;Y)}, \tag{4.13}$$

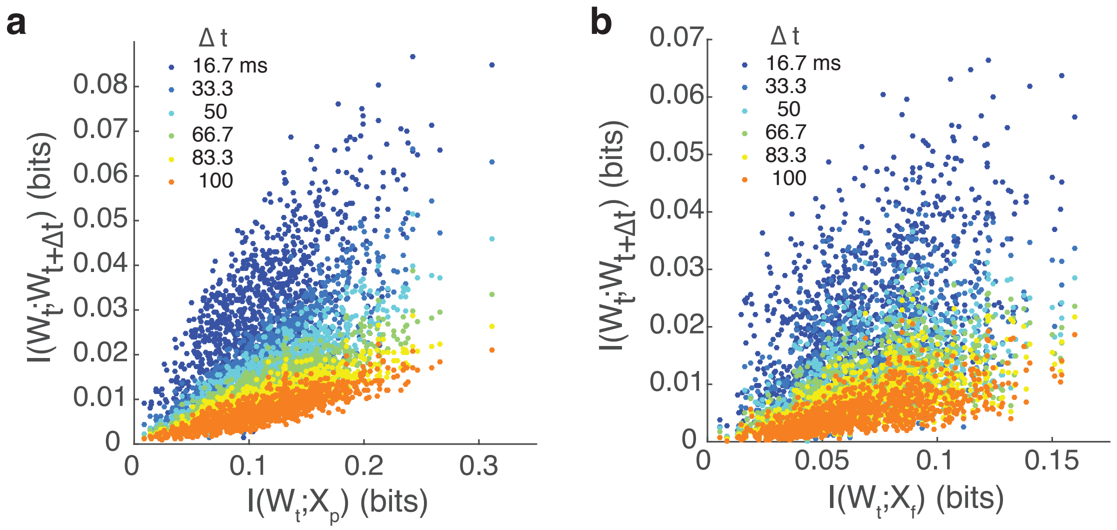

Figure 4.4: **Autoinformation and stimulus information. a.** The autoinformation at different delays plotted against stimulus information for the natural movie experiment. **b.** The autoinformation at different delays plotted against I^{future} for the moving bar experiment in Chapter 2 (statistics 2).

where $I(R^*; Y)$ is a function of the stimulus information $I(R; X)$.

We find that classical efficiency values for natural movie stimuli are modest (0.32 ± 0.07 SD) (**Figure 4.5a**). Once again, we cannot measure the information bottleneck efficiency for natural movies, so we return to the moving bar experiment. There, we find even smaller values for the classical efficiency (0.10 ± 0.05 SD), and substantially higher values for the information bottleneck efficiency (0.52 ± 0.07 SD) (**Figure 4.5b**)

4.4 Methods

4.4.1 Experimental protocol

The population recording of retinal ganglion cells responding to natural movies was collected by Stephanie Palmer and Olivier Marre in the lab of Michael Berry at Princeton University. The procedure was similar to that described in Chapter 3, using larval tiger salamanders. See [42] for further details.

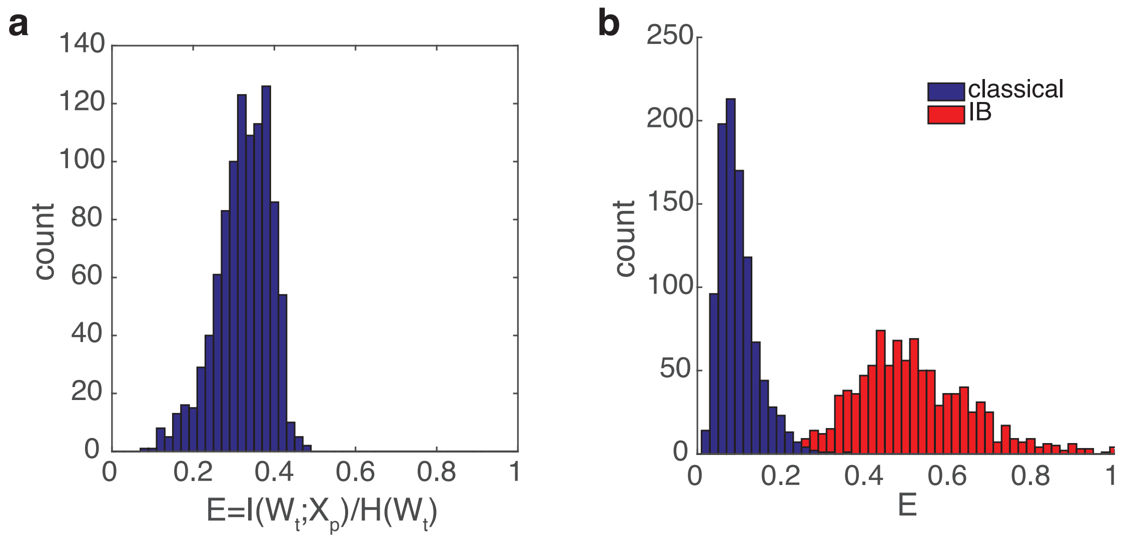

Figure 4.5: **Efficiency histograms. a.** Histogram of the classical efficiency for 1000 5 cell groups for the natural movie experiment. **b.** Histograms of the classical and information bottleneck efficiency for 1000 5 cell groups for the moving bar experiment.

4.4.2 Natural movie stimuli

Natural movies had a framerate of 60 Hz and were all 20 s long, with the exception of the 'branches' movie, which was 10 s long and repeated twice for each trial. All movies were recorded with a fixed camera with the exception of the 'camera motion' movie. Movies were presented in pseudorandom order. There were a total of about 90 trials for each stimulus.

4.5 Discussion

Our results are consistent with the idea that the retina extracts predictive information from the stimulus, and that this leads to predictability in the neural responses. However, this may not be the only reason to find some correlation in the responses. In the scenario where input noise is high (that is, the input to R is contaminated with noise in addition to the output) the optimal efficient coding filter should do some smoothing in order to mitigate the effects of noise [56, 18].

The fact that the internal predictive information is highly correlated with information about the future suggests that prediction is indeed at odds with the decorrelation of neural responses expected from efficient coding. Our formulation of the information bottleneck problem captures precisely this trade-off between efficiency and prediction. The resulting optimal filter has a nice interpretation in terms of this trade-off: The first term decorrelates the stimulus and hence leads to efficiency, while the second amplifies predictive components by an amount controlled by the trade-off parameter β.

Overall, the results in this chapter are much more indirect than those presented in the previous chapter, owing to the fact that prediction of arbitrary natural movies is a much more difficult problem than prediction of the one dimensional motion of a moving bar. We argue only that our results are consistent with our theory without constituting direct proof. Future theoretical work will be required to sharpen the predictions of our theory so that it can be more rigorously compared to efficient coding using experimental data.

CHAPTER 5

DISCUSSION

Is the retina optimized for prediction? The most direct evidence comes from Chapter 3, in which we actually measure the information the retina contains about the future trajectory of a simplified stimulus and compare this to the theoretical optimum. The results depend strongly on the statistics of the stimulus, which can be summarized by the velocity autocorrelation function. When the correlation time is short (less than 100 ms) prediction is far from optimal, and the retina can only tell us about the past stimulus, whereas when the correlation time is long (about 300 ms), prediction is very close to optimal. These results make sense in light of the empirical velocity autocorrelation functions we measure in Chapter 2. There we find velocity correlation times of at least 100 ms, with many subjects falling in the range of several hundred milliseconds. This timescale of several hundred milliseconds appears to be very important in the context of prediction, since this is also the timescale on which responses to natural stimuli are correlated, as shown in Chapter 4. While this does not constitute direct proof of optimal prediction of natural stimuli, it is consistent with our hypothesis, whereas efficient coding suggests that responses should be as decorrelated as possible.

Future work will attempt to find what range of stimulus parameters the retina is optimized for with finer granularity, in order to determine exactly where optimal prediction breaks down. Comparison of our results with [42] suggests an interesting species difference, in which the salamander retina is tuned to motion on a significantly shorter timescale. A more direct comparison, showing identical stimuli at a range of timescales to the two retinas, will be needed. We could also attempt to collect and analyze more ethologically relevant natural movies for the salamander and rat, respectively, in order to determine the source of this potential species difference.

We have focused here on prediction based on a particularly simple kind of motion—the translation of objects. While it is true that, locally, any kind of motion can be approximated by translation, more complex motions, such as the expansion and contraction caused by movement through depth, could be identified on a larger spatial scale and used as a powerful tool for prediction. This would require a more complex model for motion that may be less applicable to the retina, but selectivity for complex motion has been found, for example, in the medial superior temporal area of primate cortex [34].

One approach to making more complex predictions is to take advantage of machine learning techniques. A neural network could be trained to make predictions with a large database of natural movies. The reduced representation that it learns could then be used to estimate the amount of predictive information in a natural stimulus, which could then be compared to neural recordings.

Predicting the future is certainly not the only task of the visual system. The immense diversity of ganglion cell types in the retina [5] speaks to the range of computational functions it can perform, and it is worth investigating other choices of relevance variables in the same framework introduced in Chapter 1. Prediction may be one of the most fundamental tasks sensory systems perform, and the results presented here suggest that it should be taken seriously as a design principle for the retina.

APPENDIX A
THE FREQUENCY DOMAIN INFORMATION BOTTLENECK

Here we solve a constrained version of the information bottleneck (IB) problem [53] in the frequency domain. Let $x(t), y(t)$ be two zero-mean Gaussian processes with power spectra $S_X(f), S_Y(f)$ and coherence $\gamma(f)$. The goal of the IB problem is to find a compressed representation $r(t)$ of $x(t)$ which retains as much information as possible about $y(t)$. We restrict ourselves to linear mappings with additive Gaussian noise, $n(t)$ with spectrum $S_N(f)$, i.e.,

$$r(t) \;=\; (a * x)(t) + n(t), \tag{A.1}$$

where $a(t)$ is the impulse response of the encoding filter and $*$ denotes convolution. The power spectrum $S_R(f)$ of the compressed representation is just

$$S_R(f) \;=\; |A(f)|^2 S_X(f) + S_N(f). \tag{A.2}$$

We would like to find the filter $a(t)$ which minimizes the information theoretic cost function

$$\mathcal{L} \;=\; I(x;r) - \beta I(r;y). \tag{A.3}$$

To gain some intuition about the results, let's first solve the simpler problem of maximizing the information r contains about x while minimizing the entropy of the r, i.e.,

$$\min_a \mathcal{L} \;= H(r) - \beta I(r;x). \tag{A.4}$$

This is a formulation of the classic efficient coding hypothesis that highlights its similarity to the IB problem. It is a simplified version of a problem solved by van Hateren [56], who used a more realistic encoding model with multiple filtering stages and noise sources. Since the entropy of a continuous variable can be difficult to work with, we use the total power of r as a proxy for $H(r)$, as in [56], and use the formula for information in a Gaussian channel [12],

$$I(X;R) = -\int \log(1 - \gamma_{XR}(f))df, \tag{A.5}$$

where

$$\begin{aligned} \gamma_{XR} &= \frac{|S_{XR}|^2}{S_X S_R} && \text{(A.6)} \\ &= \frac{|A|^2 S_X}{S_R} && \text{(A.7)} \end{aligned}$$

is the coherence of x and r ($S_{XR} = |A|^2 S_X$ is the cross-spectrum of x and r). The minimization problem becomes

$$\min_{|A|^2} \mathcal{L} = \int S_R df + \beta \int \log \frac{S_N}{S_R} df. \tag{A.8}$$

Taking the derivative and setting it to zero, we obtain

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial |A|^2} &= \int S_X - \beta \frac{S_R}{S_N} \frac{S_N S_X}{S_R^2} df && \text{(A.9)} \\ 0 &= S_X - \beta \frac{S_X}{|A|^2 S_X + S_N} && \text{(A.10)} \\ |A(f)|^2 &= \frac{\beta - S_N}{S_X(f)}. && \text{(A.11)} \end{aligned}$$

We verify this is a minimum by showing that the second derivative is positive:

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial |A|^4} &= \beta \int \frac{S_X^2}{(|A|^2 S_X + S_N)^2} df && \text{(A.12)} \\ &> 0. && \text{(A.13)} \end{aligned}$$

Thus, the optimal filter is a *whitening* filter which perfectly decorrelates the input, leaving filter output that is constant at all frequencies:

$$
\begin{aligned}
S_R(f) &= (\frac{\beta - S_N}{S_X(f)})S_X(f) + S_N(f) \qquad &(\text{A.14})\\
&= \beta. \qquad &(\text{A.15})
\end{aligned}
$$

We now return to IB problem, A.3. Using

$$
\begin{aligned}
\gamma_{RY} &= \frac{|A|^2 \gamma S_X S_Y}{S_R S_Y} \qquad &(\text{A.16})\\
&= \frac{|A|^2 S_X \gamma}{S_R} \qquad &(\text{A.17})
\end{aligned}
$$

we have

$$
\min_{|A|^2} \mathcal{L} = -\int \log \frac{S_N}{S_R} df + \beta \int \log \frac{S_R - |A|^2 S_X \gamma}{S_R} df. \tag{A.18}
$$

We differentiate $\mathcal{L}$ and set it to zero:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial |A|^2} &= \int \frac{S_R}{S_N} \frac{S_N S_X}{S_R^2} + \beta \frac{S_R}{S_R - |A|^2 S_X \gamma} \frac{S_R S_X (1-\gamma) - S_X (S_R - |A|^2 S_X \gamma)}{S_R^2} df \qquad &(\text{A.19})\\
&= \int \frac{S_X}{S_R} + \beta (\frac{S_X(1-\gamma)}{S_R - |A|^2 S_X \gamma} - \frac{S_X}{S_R}) df \qquad &(\text{A.20})\\
0 &= (1-\beta)\frac{S_X}{S_R} + \beta \frac{S_X(1-\gamma)}{|A|^2 S_X (1-\gamma) + S_N}. \qquad &(\text{A.21})
\end{aligned}
$$

Since $0 \leq \gamma \leq 1$, $1 - \gamma \geq 0$. Hence, all of the terms except for $1 - \beta$ are non-negative, so

$\frac{\partial \mathcal{L}}{\partial |A|^2} = 0$ implies $\beta \geq 1$. Solving for $|A|^2$, we find

$$
\begin{aligned}
(\beta - 1)\frac{S_X}{S_R} &= \beta \frac{S_X(1-\gamma)}{|A|^2 S_X(1-\gamma) + S_N} && \text{(A.22)} \\
\frac{\beta}{\beta-1} S_R S_X (1-\gamma) &= S_X[|A|^2 S_X(1-\gamma) + S_N] && \text{(A.23)} \\
\frac{\beta}{\beta-1}(|A|^2 S_X + S_N)(1-\gamma) &= |A|^2 S_X(1-\gamma) + S_N && \text{(A.24)} \\
(\frac{\beta}{\beta-1} - 1)(1-\gamma)|A|^2 S_X &= S_N[1 - \frac{\beta}{\beta-1}(1-\gamma)] && \text{(A.25)} \\
|A|^2 &= \frac{S_N}{S_X} \frac{1 - \frac{\beta}{\beta-1}(1-\gamma)}{(\frac{\beta}{\beta-1} - 1)(1-\gamma)} && \text{(A.26)} \\
|A|^2 &= \frac{S_N}{S_X} \frac{\frac{1}{1-\gamma} - \frac{\beta}{\beta-1}}{\frac{\beta}{\beta-1} - 1} && \text{(A.27)} \\
|A|^2 &= \frac{S_N}{S_X} \frac{\frac{\beta-1}{1-\gamma} - \beta}{\beta - (\beta-1)} && \text{(A.28)} \\
|A|^2 &= \frac{S_N}{S_X} \frac{\beta - 1 - \beta(1-\gamma)}{1-\gamma} && \text{(A.29)} \\
|A(f)|^2 &= \frac{S_N}{S_X(f)} \frac{\beta\gamma(f) - 1}{1-\gamma(f)} && \text{(A.30)}
\end{aligned}
$$

Thus the optimal filter that solves the IB problem is the product (in the frequency domain) of a whitening filter and a second *relevance* filter that is a function of the coherence. Specifically, the relevance filter amplifies the informative frequencies and attenuates the uninformative frequencies. For $|A|^2$ to have a positive real solution at a given frequency, we require $\beta > \gamma^{-1}(f)$; otherwise that frequency should be filtered out completely. This is analogous to the structural phase transitions encountered in the non-dynamical multidimensional Gaussian case [17]: Increasing β allows us to smoothly capture more and more information about the relevance variable, starting with the most informative frequencies and working our way downward. Notably, the response spectrum is no longer white—it is a monotonically increasing function of the coherence, subject to a threshold set by β:

$$S_R(f) \quad = \quad \begin{cases} S_N \dfrac{\beta\gamma(f) - 1}{1 - \gamma(f)} + S_N & \text{if } \gamma(f) > \beta^{-1}, \\ S_N & \text{if } \gamma(f) \leq \beta^{-1}. \end{cases} \tag{A.31}$$

APPENDIX B

THE AUTOINFORMATION OF A NEURAL POPULATION

A great deal of theoretical and experimental neuroscience research has emphasized the importance of predictive neural computation. Central to the problem of prediction is identifying the temporal structure in the sequence of action potentials that constitute the input to a given downstream population. The most general measure of this structure is the mutual information between population spiking patterns in two non-overlapping bins. Here we study the behavior of the mutual information in the limit of small bin sizes, assuming that spiking is an inhomogeneous Poisson process with correlated rates.

Let $\mathbf{r}(t) = \begin{bmatrix} r_1(t) & \dots & r_k(t) \end{bmatrix}^\mathsf{T}$ denote the instantanous firing rates of the population of k neurons, and let $\mathbf{n}(t) = \begin{bmatrix} n_1(t) & \dots & n_k(t) \end{bmatrix}^\mathsf{T}$ be their spike counts in the interval $[t - \Delta t/2, t + \Delta t/2]$. We assume the firing rates are stationary and have (unnormalized) auto- and cross-correlation functions given by

$$\rho_{ij}(\tau) = \mathbf{E}[r_i(t) r_j(t+\tau)]. \tag{B.1}$$

Note that these correlation functions are identical to the limiting values of the binned spike train correlation functions, normalized by Δt^2 (Theorem 1 in [28]),

$$\lim_{\Delta t \to 0} \frac{\mathbf{E}[n_i(t) n_j(t+\tau)]}{\Delta t^2} = \rho_{ij}(\tau) \tag{B.2}$$

(except where the limit diverges at $\tau = 0$ for $i = j$), which provides a means of estimating them from data. The normalization by Δt^2 is necessary because $\lim_{\Delta t \to 0} \mathbf{E}[n_i(t) n_j(t+\tau)] = 0$. The same is true of the mutual information; thus, we are interested in examining the behavior of the following function, which we call the autoinformation:

$$\mathcal{I}(\tau) = \lim_{\Delta t \to 0} \frac{I(\mathbf{N}(t); \mathbf{N}(t+\tau))}{\Delta t^2} \tag{B.3}$$

where

$$I(\mathbf{N}(t);\mathbf{N}(t+\tau)) = \sum_{\mathbf{n}(t),\mathbf{n}(t+\tau)} p(\mathbf{n}(t),\mathbf{n}(t+\tau)) \log \frac{p(\mathbf{n}(t),\mathbf{n}(t+\tau))}{p(\mathbf{n}(t))p(\mathbf{n}(t+\tau))}. \tag{B.4}$$

Our approach to solving (B.3) is similar to that of [43]. Note that in [43], the authors find a second order expansion of the mutual information that holds for sufficiently small Δt; here we equivalently find the limit (B.3), from which we can recover the mutual information (B.4) at sufficiently small Δt simply by multiplying by Δt^2. When Δt is small, $N_i(t)$ will be approximately a Bernoulli distributed binary variable with $p(N_i(t) = 1) = r_i(t)\Delta t$ and $p(N_i(t) = 0) = 1 - r_i(t)\Delta t$, which we will use to calculate the distributions in terms of the rates, $\mathbf{r}$. We denote the binary vector in which a subset of neurons $S \subseteq K = \{1,\dots,k\}$ spikes and all others are silent by $\mathbf{e}_S$. The joint probability of two such vectors is

$$\begin{aligned} p(\mathbf{e}_{S_1},\mathbf{e}_{S_2}) = \dots \\ \mathbf{E}[\prod_{i\in S_1} r_i(t)\Delta t \prod_{i\in K\setminus S_1}(1-r_i(t)\Delta t) \prod_{i\in S_2} r_i(t+\tau)\Delta t \prod_{i\in K\setminus S_2}(1-r_i(t+\tau)\Delta t)]. \end{aligned} \tag{B.5}$$

For $s = |S_1| + |S_2|$ total spikes, the joint distribution is a sum of terms scaling with power of Δt, from Δt^s to Δt^k. When $s > 2$, $\lim_{\Delta t\to 0} p(\mathbf{e}_{S_1},\mathbf{e}_{S_2})/\Delta t^2 = 0$. Thus, the only response patterns that will contribute to the autoinformation are those with zero, one, or two spikes. We denote these patterns $\mathbf{0} = \mathbf{e}_\emptyset$, $\mathbf{e}_i = \mathbf{e}_{\{i\}}$, and $\mathbf{e}_{ij} = \mathbf{e}_{\{i,j\}}$. We see that the two spike patterns also do not contribute: using $\mathcal{O}(x)$ to denote terms of order x and smaller, the probabilities are

$$p(\mathbf{e}_{ij},\mathbf{0}) = \rho_{ij}(\tau)\Delta t^2 + \mathcal{O}(\Delta t^3), \tag{B.6}$$

$$p(\mathbf{e}_{ij})p(\mathbf{0}) = \rho_{ij}(\tau)\Delta t^2 + \mathcal{O}(\Delta t^3), \tag{B.7}$$

so

$$\lim_{\Delta t\to 0} \frac{p(\mathbf{e}_{ij},\mathbf{0})}{\Delta t^2} \log \frac{p(\mathbf{e}_{ij},\mathbf{0})}{p(\mathbf{e}_{ij})p(\mathbf{0})} = 0. \tag{B.8}$$

The remaining terms of the autoinformation converge to finite values as follows. For the case of zero spikes in both bins we have

$$\begin{aligned} p(\mathbf{0},\mathbf{0}) &= 1-2\sum_i \bar{r}_i\Delta t + 2\sum_{i<j}\rho_{ij}(0)\Delta t^2 + \sum_{i,j}\rho_{ij}(\tau)\Delta t^2 + \mathcal{O}(\Delta t^3), && \text{(B.9)} \\ p(\mathbf{0})p(\mathbf{0}) &= 1-2\sum_i \bar{r}_i\Delta t + 2\sum_{i<j}\rho_{ij}(0)\Delta t^2 + \sum_{i,j}\bar{r}_i\bar{r}_j\Delta t^2 + \mathcal{O}(\Delta t^3), && \text{(B.10)} \end{aligned}$$

where $\bar{r}_i = \mathbf{E}[r_i]$. To calculate the limit, we begin by using the approximation

$$\log(1-x) \approx -x. \tag{B.11}$$

which holds for small x. Here,

$$x = 1 - \frac{p(\mathbf{0},\mathbf{0})}{p(\mathbf{0})p(\mathbf{0})} = \frac{\sum_{i,j}[\bar{r}_i\bar{r}_j - \rho_{ij}(\tau)]\Delta t^2}{p(\mathbf{0})p(\mathbf{0})} + \mathcal{O}(\Delta t^3) \tag{B.12}$$

which leads to

$$\lim_{\Delta t\to 0}\frac{p(\mathbf{0},\mathbf{0})}{\Delta t^2}\log\frac{p(\mathbf{0},\mathbf{0})}{p(\mathbf{0})p(\mathbf{0})} = \sum_{i,j}[\rho_{ij}(\tau) - \bar{r}_i\bar{r}_j]. \tag{B.13}$$

For a spike in the first time bin and not the second, we have

$$\begin{aligned} p(\mathbf{e}_i,\mathbf{0}) &= \bar{r}_i\Delta t - \sum_{j\neq i}\rho_{ij}(0)\Delta t^2 - \sum_j \rho_{ij}(\tau)\Delta t^2 + \mathcal{O}(\Delta t^3), && \text{(B.14)} \\ p(\mathbf{e}_i)p(\mathbf{0}) &= \bar{r}_i\Delta t - \sum_{j\neq i}\rho_{ij}(0)\Delta t^2 - \sum_j \bar{r}_i\bar{r}_j\Delta t^2 + \mathcal{O}(\Delta t^3), && \text{(B.15)} \end{aligned}$$

which, by the same procedure, yields

$$\lim_{\Delta t\to 0}\frac{p(\mathbf{e}_i,\mathbf{0})}{\Delta t^2}\log\frac{p(\mathbf{e}_i,\mathbf{0})}{p(\mathbf{e}_i)p(\mathbf{0})} = \sum_j[\bar{r}_i\bar{r}_j - \rho_{ij}(\tau)]. \tag{B.16}$$

The same equation holds for a spike in the second time bin and not the first. Finally, for one spike in each time bin,

$$\begin{aligned} p(\mathbf{e}_i, \mathbf{e}_j) &= \rho_{ij}(\tau)\Delta t^2 + \mathcal{O}(\Delta t^3), & \text{(B.17)} \\ p(\mathbf{e}_i)p(\mathbf{e}_j) &= \bar{r}_i\bar{r}_j\Delta t^2 + \mathcal{O}(\Delta t^3), & \text{(B.18)} \end{aligned}$$

and calculation of the limit is straightforward:

$$\lim_{\Delta t \to 0} \frac{p(\mathbf{e}_i, \mathbf{e}_j)}{\Delta t^2} \log \frac{p(\mathbf{e}_i, \mathbf{e}_j)}{p(\mathbf{e}_i)p(\mathbf{e}_j)} = \rho_{ij}(\tau) \log \frac{\rho_{ij}(\tau)}{\bar{r}_i\bar{r}_j}. \tag{B.19}$$

Finally, to calculate B.3, we must take the sum over all the contributing patterns, i.e., B.13, plus two copies of B.16, summed over neurons i, plus B.19, summed over all combinations of neurons i, j. All of the terms are double sums over i and j, yielding

$$\mathcal{I}(\tau) = \sum_{i,j} \mathcal{I}_{ij}(\tau). \tag{B.20}$$

where

$$\mathcal{I}_{ij}(\tau) = \rho_{ij}(\tau) \log \frac{\rho_{ij}(\tau)}{\bar{r}_i\bar{r}_j} - (\rho_{ij}(\tau) - \bar{r}_i\bar{r}_j). \tag{B.21}$$

After some rewriting, we see that each cell contributes its own autoinformation plus the information it carries about each other cell in the population:

$$\mathcal{I}(\tau) = \sum_i [\mathcal{I}_{ii}(\tau) + \sum_{j \neq i} \mathcal{I}_{ij}(\tau)]. \tag{B.22}$$

The fact that the total autoinformation decomposes this way tells us something about the redundancy of the population autoinformation in this limit. Redundancy is the difference between the sum of the information that the individual cells carry about the population and the total autoinformation, but as we see above, these two quantities are identical, so the redundancy must go to zero in the limit.

REFERENCES

[1] Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[2] Evan Archer, Il Memming Park, and Jonathan W Pillow. Bayesian and quasi-bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755, 2013.

[3] Evan W Archer, Il Memming Park, and Jonathan W Pillow. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. In *Advances in Neural Information Processing Systems*, pages 1700–1708, 2013.

[4] Stephen A Baccus, Bence P Ölveczky, Mihai Manu, and Markus Meister. A retinal circuit that computes object motion. *Journal of Neuroscience*, 28(27):6807–6817, 2008.

[5] Tom Baden, Philipp Berens, Katrin Franke, Miroslav Román Rosón, Matthias Bethge, and Thomas Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345, 2016.

[6] H. B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[7] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711, 2012.

[8] Michael J II Berry, Iman H Brivanlou, Thomas A Jordan, and Markus Meister. Anticipation of moving stimuli by the retina. *Nature*, 398(6725):334, 1999.

[9] William Bialek, Rob R de Ruyter van Steveninck, and Naftali Tishby. Efficient representation as a design principle for neural coding and computation. In *Information Theory, 2006 IEEE International Symposium on*, pages 659–663. IEEE, 2006.

[10] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

[11] Bart G Borghuis and Anthony Leonardo. The role of motion extrapolation in amphibian prey capture. *Journal of Neuroscience*, 35(46):15430–15441, 2015.

[12] Alexander Borst and Frédéric E Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947, 1999.

[13] Clifford P Brangwynne, Gijsje H Koenderink, Frederick C MacKintosh, and David A Weitz. Intracellular transport by active diffusion. *Trends in Cell Biology*, 19(9):423–427, 2009.

[14] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European Conference on Computer Vision*, pages 282–295. Springer, 2010.

[15] Dirk Calow and Markus Lappe. Local statistics of retinal optic flow for self-motion through natural sceneries. *Network: Computation in Neural Systems*, 18(4):343–374, 2007.

[16] Avi Caspi, Rony Granek, and Michael Elbaum. Enhanced diffusion in active intracellular transport. *Physical Review Letters*, 85(26):5655, 2000.

[17] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

[18] Eizaburo Doi and Michael S Lewicki. A simple model of optimal population coding for sensory systems. *PLoS Computational Biology*, 10(8):e1003761, 2014.

[19] Peter Elias. Predictive coding–i. *IRE Transactions on Information Theory*, 1(1):16–24, 1955.

[20] Peter Elias. Predictive coding–ii. *IRE Transactions on Information Theory*, 1(1):24–33, 1955.

[21] David J Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, 1994.

[22] Daniel T Gillespie. The mathematics of brownian motion and johnson noise. *American Journal of Physics*, 64(3):225–240, 1996.

[23] Ido Golding and Edward C Cox. Physical nature of bacterial cytoplasm. *Physical Review Letters*, 96(9):098102, 2006.

[24] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[25] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[26] Dirk Jancke, Wolfram Erlhagen, Gregor Schöner, and Hubert R Dinse. Shorter latencies for motion trajectories than for flashes in population responses of cat primary visual cortex. *The Journal of Physiology*, 556(3):971–982, 2004.

[27] Jamie Johnston and Leon Lagnado. General features of the retinal connectome determine the computation of motion anticipation. *eLife*, 4, 2015.

[28] Michael Krumin and Shy Shoham. Generation of spike trains with controlled auto-and cross-correlation functions. *Neural Computation*, 21(6):1642–1664, 2009.

[29] George V Lauder and Peter GA Madden. Advances in comparative physiology from high-speed imaging of animal and fluid motion. *Annu. Rev. Physiol.*, 70:143–163, 2008.

[30] Anthony Leonardo and Markus Meister. Nonlinear dynamics support a linear population code in a retinal target-tracking circuit. *Journal of Neuroscience*, 33(43):16971–16982, 2013.

[31] Michael S Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356, 2002.

[32] Olivier Marre, Dario Amodei, Nikhil Deshmukh, Kolia Sadeghi, Frederick Soo, Timothy E Holy, and Michael J Berry. Mapping a complete neural population in the retina. *Journal of Neuroscience*, 32(43):14859–14873, 2012.

[33] Olivier Marre, Vicente Botella-Soler, Kristina D Simmons, Thierry Mora, Gašper Tkačik, and Michael J Berry II. High accuracy decoding of dynamical motion from a large retinal population. *PLoS Computational Biology*, 11(7):e1004304, 2015.

[34] Patrick J Mineault, Farhan A Khawaja, Daniel A Butts, and Christopher C Pack. Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proceedings of the National Academy of Sciences*, 109(16):E972–E980, 2012.

[35] Matteo Mischiati, Huai-Ti Lin, Paul Herold, Elliot Imler, Robert Olberg, and Anthony Leonardo. Internal models direct dragonfly interception steering. *Nature*, 517(7534):333, 2015.

[36] Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*, pages 471–478, 2002.

[37] Romi Nijhawan. Motion extrapolation in catching. *Nature*, 1994.

[38] Romi Nijhawan. Visual prediction: Psychophysics and neurophysiology of compensation for time delays. *Behavioral and Brain Sciences*, 31(2):179–198, 2008.

[39] Simon F Nørrelykke and Henrik Flyvbjerg. Harmonic oscillator in heat bath: Exact simulation of time-lapse-recorded data and exact analytical benchmark statistics. *Physical Review E*, 83(4):041103, 2011.

[40] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

[41] Bence P Ölveczky, Stephen A Baccus, and Markus Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003.

[42] Stephanie E Palmer, Olivier Marre, Michael J Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

[43] Stefano Panzeri, Simon R Schultz, Alessandro Treves, and Edmund T Rolls. Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1423):1001–1012, 1999.

[44] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79, 1999.

[45] Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: Exploring the Neural Code*. MIT Press, 1999.

[46] Stefan Roth and Michael J Black. On the spatial statistics of optical flow. *International Journal of Computer Vision*, 74(1):33–50, 2007.

[47] Greg Schwartz, Rob Harris, David Shrom, and Michael J Berry II. Detection and prediction of periodic patterns by the retina. *Nature Neuroscience*, 10(5):552, 2007.

[48] Greg Schwartz, Sam Taylor, Clark Fisher, Rob Harris, and Michael J Berry II. Synchronized firing among retinal ganglion cells signals motion reversal. *Neuron*, 55(6):958–969, 2007.

[49] Claude E Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:623–656, 1948.

[50] Mandyam V Srinivasan, Simon B Laughlin, and Andreas Dubs. Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London B: Biological Sciences*, 216(1205):427–459, 1982.

[51] Manivannan Subramaniyan, Alexander S Ecker, Saumil S Patel, James R Cotton, Matthias Bethge, Philipp Berens, and Andreas S Tolias. Faster processing of moving compared to flashed bars in awake macaque v1 provides a neural correlate of the flash lag illusion. *bioRxiv*, page 031146, 2015.

[52] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2432–2439. IEEE, 2010.

[53] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[54] Stuart Trenholm, David J Schwab, Vijay Balasubramanian, and Gautam B Awatramani. Lag normalization in an electrically coupled neural network. *Nature Neuroscience*, 16(2):154, 2013.

[55] J Hans van Hateren and Dan L Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1412):2315–2320, 1998.

[56] Johannes H van Hateren. A theory of maximizing sensory information. *Biological Cybernetics*, 68(1):23–29, 1992.

[57] Henning U Voss. The leaky integrator with recurrent inhibition as a predictor. *Neural Computation*, 28(8):1498–1502, 2016.

[58] Henning U Voss. Signal prediction by anticipatory relaxation dynamics. *Physical Review E*, 93(3):030201, 2016.

[59] David K Warland, Pamela Reinagel, and Markus Meister. Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78(5):2336–2350, 1997.

[60] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13, 2006.