Quantum of Vision

Thesis by
Bo Chen

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2016
Defended May 12, 2016

Bo Chen
ORCID: 000-0001-5566-736

ACKNOWLEDGEMENTS

I enjoy doing research that is different. Being different means not inheriting the definitions and solutions of existing problems but striving to recognize and address new problems. This notion of being different is imparted to me by my advisor Prof. Pietro Perona, to whom I am deeply grateful. I have consulted Pietro for basically everything: the meaning of life, the purpose of research, strategies for picking research projects, strategies for picking engagement rings, the best wedding venues, and the best obstetricians. Thank you, Pietro, for being the walking embodiment of Google and what Google could only aspire to become. I would also like to express my heartfelt gratitude towards Prof. Ueli Rutishauser and Dr. Qi Zhao for indulging my interdisciplinary wanderlust, as well as Dr. Lubomir Bourdev and Dr. Yang Song for connecting my research with practice. Furthermore, my sincere appreciation goes to my thesis committee, including Prof. Andreas Krause, Prof. Markus Meister, Dr. Victoria Kostina, Prof. Doris Tsao and Prof. Colin Camerer, who have become an integral part of my scholarly upbringing. I will also miss the thoughtful dinner conversations with Dr. Steve Branson, Dr. Michael Maire, and Dr. Dan McNamee.

Being different also comes with a cost, as the road less traveled is often traveled alone. The past six years have been an uphill battle where I have constantly been on the losing side. The intellectual solitude was bearable owing to the social companionship of my dear family and friends. I am blessed with the great friend Linsanity Yongjun, who was always there when I needed a morale boost. I am indebted to all members of the Vision lab for their moral support, with a special hat tip to Krzysztof Chalupka, Matteo Ruggero Ronchi and Joe Marino for going out of their way to ensure my wellbeing. Dr. Tatiana Vasilevskaia and the Caltech counseling center were also tremendously helpful. Just as this thesis will discuss how vision systems can see in the dark with only a few particles of light, I saw the way forward in my darkest moments thanks to the small glimpses of hope: every kind word and every pat on the shoulder brought me closer to the finishing line. The brightest beacon of hope was my wife Kaida, who supported my journey with unconditional optimism, unique perspectives and unquestionably exquisite culinary skills, and to whom I wholeheartedly devote this thesis.

ABSTRACT

Visual inputs to artificial and biological visual systems are often quantized: cameras accumulate photons from the visual world, and the brain receives action potentials from visual sensory neurons. Collecting more information quanta leads to a longer acquisition time and better performance. In many visual tasks, collecting a small number of quanta is sufficient to solve the task well. The ability to determine the right number of quanta is pivotal in situations where visual information is costly to obtain, such as photon-starved or time-critical environments. In these situations, conventional vision systems that always collect a fixed and large amount of information are infeasible. I develop a framework that judiciously determines the number of information quanta to observe based on the cost of observation and the requirement for accuracy. The framework implements the optimal speed versus accuracy tradeoff when two assumptions are met, namely that the task is fully specified probabilistically and constant over time. I also extend the framework to address scenarios that violate the assumptions. I deploy the framework to three recognition tasks: visual search (where both assumptions are satisfied), scotopic visual recognition (where the model is not specified), and visual discrimination with unknown stimulus onset (where the model is dynamic over time). Scotopic classification experiments suggest that the framework leads to dramatic improvement in photon-efficiency compared to conventional computer vision algorithms. Human psychophysics experiments confirmed that the framework provides a parsimonious and versatile explanation for human behavior under time pressure in both static and dynamic environments.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] B. Chen and P. Perona, "Vision without the image," *Sensors*, vol. 16, no. 4, pp. 484–484, 2016.

[2] ——, "Scotopic visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015, pp. 8–11.

[3] ——, "Speed versus accuracy in visual search: Optimal performance and neural architecture," *Journal of Vision*, vol. 15, no. 16, pp. 9–9, 2015.

[4] B. Chen, P. Perona, and L. D. Bourdev, "Hierarchical cascade of classifiers for efficient poselet evaluation.," in *British Machine Vision Conference (BMVC)*, 2014.

[5] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[6] B. Chen, A. Krause, and R. M. Castro, "Joint optimization and variable selection of high-dimensional Gaussian processes," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012, pp. 1423–1430.

[7] B. Chen, V. Navalpakkam, and P. Perona, "Predicting response time and error rates in visual search," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Quantized visual information

Images are the dominant medium through which we make sense of the world. Computer vision systems analyze images to extract information about the environment (e.g. understanding the identities of and relationships between people in a meeting room); neuroscientists and psychophysicists study the primate vision system using image stimuli (e.g. study human gaze patterns in response to an image of the beach). The use of images divides visual perception into two stages: information acquisition (forming the image) and analysis (understanding what is inside the image).

We study a different type of vision system where information acquisition and analysis are not divided but intertwined. These vision systems collect visual information one small quantum at a time, and analyze the quanta as they arrive. For example, a camera senses photons from the surrounding environment. Every photon falling on a particular pixel contains information about the visual area corresponding to the pixel, and thus can update the vision system's belief about what is in the environment. The photon is thus an indivisible piece of visual information, which we refer to as a "visual quantum". The use of visual quanta as an alternative medium to images may be justified in the following examples.

First, acquiring images may be quite expensive in low light environments, and the long exposure is often undesirable: in biological imaging, prolonged exposure could cause health risks [1] or sample bleaching [2]; in autonomous driving, the delay imposed by image capture could affect a vehicle's ability to stay on-course and avoid obstacles [3]; in surveillance, long periods of imaging could delay response, produce smeared images, or compromise stealth. In these scenarios, instead of waiting for a high-quality image after a long exposure, visual systems should process every single photon as it arrives, and make a decision as soon as sufficient photons have been collected.

Second, the quantized view is consistent with the information processing mechanism of biological visual systems. To transmit information from one area to the next (e.g. from the retina to the visual cortex), the visual system uses action potentials or "spikes" [4]. Action potentials, like the photons, are quantized: the impulses have a

stereotypical shape, and information resides in the timing and the counts. Similarly, the quantization becomes useful when time is critical. When the visual system is under time-pressure (e.g. search for predator or prey), it must exploit every single action potential to make a decision as quickly and as accurately as possible [5]. Hence modeling the quantized signal may help neuroscientists and psychophysicists understand visual perception in humans and other animals.

Lastly, the quantized reasoning is consistent with the trend of development in sensor technology. Next-generation visual sensors will be equipped with photon-counting capabilities. For example, the Quanta Image Sensor [6] and the Giga-vision sensors [7] will detect and report single photon arrival events. The original goal of designing photon-counting sensors was to increase the signal-to-noise ratio as well as the spatial and temporal resolution for imaging. Serendipitously, the photon-counting capability also enabled vision applications to sense and compute with quantized visual information.

Moreover, quantization does not stop at the level of the sensory input – the entire computation pipeline from sensory inputs to a decision may be quantized as well. It is the case for biological visual systems, where quantized communication in the form of action potentials occur throughout all stages of computation. The quantization of the thought process may then aid neuroscientists in understanding the functional roles played by different components in the system. It is also sensible for computer vision systems to discretize computation. Since the input signals are quantized, the changes in the internal states of the system should be discretized. When the changes are sparse, a discrete implementation may be more efficient than a continuous implementation in terms of the computation time, communication cost, and energy consumption. This observation has become more relevant recently thanks to the return of artificial neural networks as the workhorse for visual recognition tasks, for which the changes are sparse and the energy is key in low light environments.

1.2 The speed versus accuracy tradeoff

Information about the world trickles in one quantum (photon, action potential, etc) at a time. It is up to the observer to decide how many quanta to collect. Collecting more information requires time while collecting too little information subjects the observer to errors. The key is to collect just the right amount of information while maintaining certain accuracy guarantees (see **Fig. 1.1** for illustration). The balance between the amount of information and the quality of the decision is called the speed

versus accuracy tradeoff (SAT).

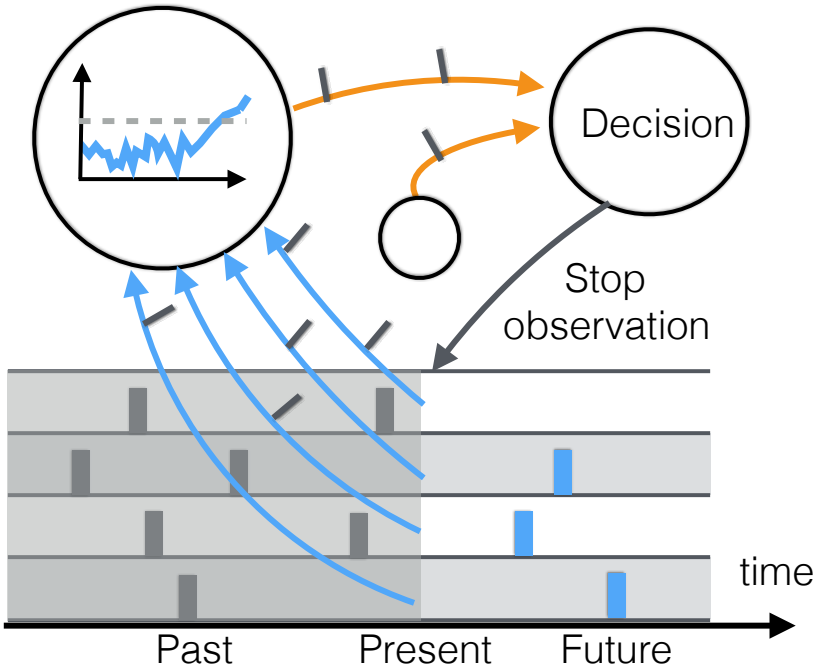

Figure 1.1: **Quantized vision**. Information trickles into a vision system (through blue arrows) one quantum at a time. The vision system may also be quantized in that computation flows through the system in small packets (through orange arrows). The quantization in the input provides flexibility to stop collecting information (through grey arrow) as soon as a decision is reached with sufficient certainty. The quantization in the internal computation provides efficiency in computation.

This thesis is about the theory and practice of SAT in visual perception tasks for biological and artificial systems. Critically, the information processing pipeline is quantized from sensory input collection to decision computation. To optimize SAT it is imperative to know how each quantum of information contributes to the task at hand, and when the cumulative information is ripe for decision. **Ch. 2** lays down the theoretical framework for answering these questions. The framework assumes that the task is fully specified by a probabilistic model that is static in time, and **Ch. 3** gives an example using visual search where both assumptions are met. In practical and ecological conditions, a probabilistic model is often not available and the vision system must learn the decision rules for optimizing SAT. Thus **Ch. 4** discusses the issue of learning with the application of visual classification in lowlight. **Ch. 5** describes a visual discrimination example where where the probabilistic model changes over time. Lastly **Ch. 6** studies the optimality of our framework in SAT,

and **Ch. 7** offers the final remarks.

> The chapters are self-contained. All readers are encouraged to start from **Ch. 2** (framework). Readers with a psychophysics and neuroscience background may read only **Ch. 3** (search) and **Ch. 5** (discrimination with unknown stimulus onset); computer vision readers may start from **Ch. 4** (classification); **Ch. 6** (optimality analysis) is reserved for the mathematically-inclined. You will find more helper texts like this that explain how to navigate the thesis and why I have done things one way instead of another.

References

[1] E. Hall and D. Brenner, "Cancer risks from diagnostic radiology," *Cancer*, vol. 81, no. 965, 2014.

[2] D. J. Stephens and V. J. Allan, "Light microscopy techniques for live cell imaging," *Science*, vol. 300, no. 5616, pp. 82–86, 2003.

[3] D. F. Llorca, V. Milanés, I. P. Alonso, M. Gavilán, I. G. Daza, J. Pérez, and M. Á. Sotelo, "Autonomous pedestrian collision avoidance using a fuzzy steering controller," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 390–401, 2011.

[4] E. R. Kandel, J. H. Schwartz, T. M. Jessell, *et al.*, *Principles of Neural Science*. McGraw-Hill New York, 2000, vol. 4.

[5] R. Vanrullen and S. J. Thorpe, "The time course of visual processing: From early perception to decision-making," *Journal of Cognitive Neuroscience*, vol. 13, no. 4, pp. 454–461, 2001.

[6] E. R. Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors," *Electron Devices Society, IEEE Journal of the*, vol. 1, no. 9, pp. 166–174, 2013.

[7] L. Sbaiz, F. Yang, E. Charbon, S. Süsstrunk, and M. Vetterli, "The gigavision camera," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, pp. 1093–1096.

Chapter 2

SEQUENTIAL PROBABILITY RATIO TEST

A framework for Analyzing Quantized Visual Input

We discuss the theoretical framework that optimizes the speed versus accuracy tradeoff (SAT) for systems with quantized visual inputs. The framework is based on a mature idea in statistics called the sequential probability ratio test (SPRT, [1]). The main goal of this chapter is to review the assumptions and optimality guarantees of SPRT.

2.1 Input assumptions

We start with the assumptions regarding the quantized inputs. These assumptions are used to develop the basic form of our framework and will be relaxed in later chapters.

Assumption 1: known probabilistic model

The first piece of the puzzle is understanding what the input is and how it is generated. Our assumption is that there exists a statistical generative model that relates the quantized inputs to important variables for solving the task at hand.

For example, when a lioness peruses a herd of buffalos on an open meadow at night, every part of visual scene conveys information – the locations of the patriarch, the calfs, the elders and the injured are useful for planning an attack. Nature communicates this information using the spatial and temporal arrangement of photons and the law of physics: the brighter a visual location is, the more photons will be reflected to hit the lioness' retina in a given amount of time. This physical law fits precisely our assumption: the inputs (photons) are generated according to physical variables (attributes of buffalo), which is useful to solve the problem (planning an attack).

This assumption also works phenomenologically: it does not require precise knowledge of the physical generative process between task-relevant properties and sensor inputs. Take a look inside the lioness' visual system. Information processing here involves neurons and action potentials, which appears completely different from the information processing that involves the retina and photons, but actually also fits the assumption. A subset of neurons in the system are selective towards elementary shapes such as edges and curves [2]. Neurons in this area will each be triggered by a

Chapter / Assumption	Known probabilistic model	Time-homogeneity
Visual search (**Ch. 3**)	✓	✓
Scotopic vision (**Ch. 4**)	×	✓
Visual discrimination (**Ch. 5**)	✓	×

Table 2.1: **The set of assumptions satisfied by each application**.

specific patch of the visual world to emit action potentials, where the emission rate reflects the shape information of the patch. If we consider the action potentials from these neurons as inputs of the visual system, it holds that the inputs (action potentials) are statistically characterized by properties of the physical world (shapes in the visual world). Therefore, despite lacking a complete understanding of the physical process of how light goes through the retina and the lateral geniculate nucleus, and then triggers the shape-selective neurons to fire (which may be quite intricate [3]), our assumption stands as long as the statistical dependency between the inputs and the properties of interest is known.

Assumption 2: time-homogeneity

Our second assumption is that the statistical model is constant over time. If both the lioness and the herd are steady enough, the photons reflected from the scene should have the same statistics regardless of how long the lioness has been scrutinizing. As a coarse approximation, the train of action potentials in the orientation-selective area of the primate visual cortex also follow the same statistics within typical durations for making a quick decision [4]. Essentially, time-homogeneity ensures that the number of observations regarding any visual property is potentially infinite, and the uncertainty around the visual property will vanish over time.

Table. 2.1 outlines the set of assumptions satisfied by the problems in each coming chapter.

2.2 Notation

Formally, the quantized inputs are the time series $\boldsymbol{X}_{1:t} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\}$, where time has been judiciously discretized into bins of size Δ, and the observation $\boldsymbol{X}_t$ spans the duration $((t-1)\Delta, t\Delta]$. Each $\boldsymbol{X}_t \in [\mathbb{Z}^+]^D$ is a D-dimensional vector. An element in $\boldsymbol{X}_t$ counts the number of visual quanta from one of D input channels. For images, $\boldsymbol{X}_t$ could be photon count and D is the number of pixels; for neurons, $\boldsymbol{X}_t$ could be spike counts and D is the number of neurons. Generally speaking, we use the subscripts to represent the channel and time i.e. $X_{i,t}$ denotes the count at neuron i and time

bin t. We also use boldface for vectors and matrices and regular font for scalars. An object of interest in the visual world (e.g. a buffalo) could be characterized as one of K categories (e.g. $K = 2$ for the categories { "weak", "strong" }). Let $C \in \{1, 2, \ldots, K\}$ denotes the category of the object. The visual system is free to report at any time t a class estimate $\hat{C}_t \in \{1, 2, \ldots, K\}$. For simplicity we only consider classification tasks: the task is to identify the class of the object, which maps one-to-one to a decision. In other words we assume that the lioness will always commit to a chase once she identifies the prey as weak, and skip the prey she deems strong.

> One might argue that identifying the category of the object and deciding on an action should be two separated tasks. For example identifying the strength of prey and deciding to give chase have different semantics. This is true, but semantic difference may be all there is. In the lioness' problem we can reformulate the categories to "attackable" and "to be avoided", and then the classification and the actions would agree.

2.3 Optimality

Now that we have specified the assumptions regarding sensory input, we are ready to define optimality. As soon as the stream of sensory input pours in, an observer faces a double decision. First, at each time instant it has to decide whether the information in the input collected so far is sufficient to reach a decision. Second, once information is deemed sufficient, it has to pick what decision to make. Moreover, the decisions must "optimally" trade off reaction time (RT), the amount of time the observer spends to collect information, with error rate (ER), the frequency of making mistakes.

Optimality is defined with respect to the Bayes risk [5], [6]:

$$\text{BayesRisk} = \mathbb{E}[T] + \eta \mathbb{E}[\hat{C}_T \neq C], \tag{2.1}$$

where $\mathbb{E}$[T] is the expected reaction time, and $\mathbb{E}[\hat{C}_T \neq C]$ is the probability of the observer committing to a wrong prediction. η is a parameter that specifies the cost of making mistakes (in seconds). For example, η might be quantified in terms of the time wasted failing to overpower a strong buffalo. The relative cost of errors and time is determined by the circumstances in which the observer operates. η may be higher if the lioness is hungry (catching the prey has higher value), or lower if the lioness is well hidden (sustained observation is more feasible).

> Why are RT and ER combined linearly in the Bayes risk? The expression originates from the general description of the observer's objective:
>
> $$\min \mathbb{E}[T], s.t. \mathbb{E}[\hat{C}_t \neq C], \leq \text{maxerr} \tag{2.2}$$
>
> where maxerr is the upper bound on the misclassification error. This constrained cost function may be concerted to an unconstrained objective via Lagrange multipliers, and the result is precisely the Bayes risk.

Thus, the Bayes risk measures the combined RT and ER costs of a given search mechanism. For now we assume that misclassification errors of different kinds all have the same cost, but this is only for simplicity and will be relaxed in future chapters.

Next we will present an efficient and popular statistical technique called the Sequential Probability Ratio Test [1] as our main algorithm for SAT optimization.

2.4 Sequential probability ratio test

SPRT is an algorithm that takes an endless streams of evidence $\boldsymbol{X}_{1:t}$ and decides (1) when to stop observing and (2) what decision to make. The classic SPRT discriminates between two classes ($K = 2$). Crucially SPRT relies on a probabilistic model that relates the class C to the observations. SPRT takes the following form (see **Fig. 2.1** for illustration):

$$S(\boldsymbol{X}_{1:t}) \triangleq \log \frac{P(C=1|\boldsymbol{X}_{1:t})}{P(C=0|\boldsymbol{X}_{1:t})} \begin{cases} \geq \tau & \text{Declare } \hat{C}_t = 1 \\ \leq -\tau & \text{Declare } \hat{C}_t = 0 \\ \text{otherwise} & t \leftarrow t+1. \end{cases} \tag{2.3}$$

It considers $S(\boldsymbol{X}_{1:t})$, the log likelihood ratio between the two classes with respect to the observations $\boldsymbol{X}_{1:t}$. The observer declares class 1 as soon as $S(\boldsymbol{X}_{1:t})$ crosses an upper threshold τ, and declares class 0 as soon as $S(\boldsymbol{X}_{1:t})$ crosses a lower threshold $-\tau$. Until either event takes place, the observer waits for further information. For convenience we use base 10 for all our logarithms and exponentials, i.e. $\log(x) \triangleq \log_{10}(x)$ and $\exp(x) \triangleq 10^x$.

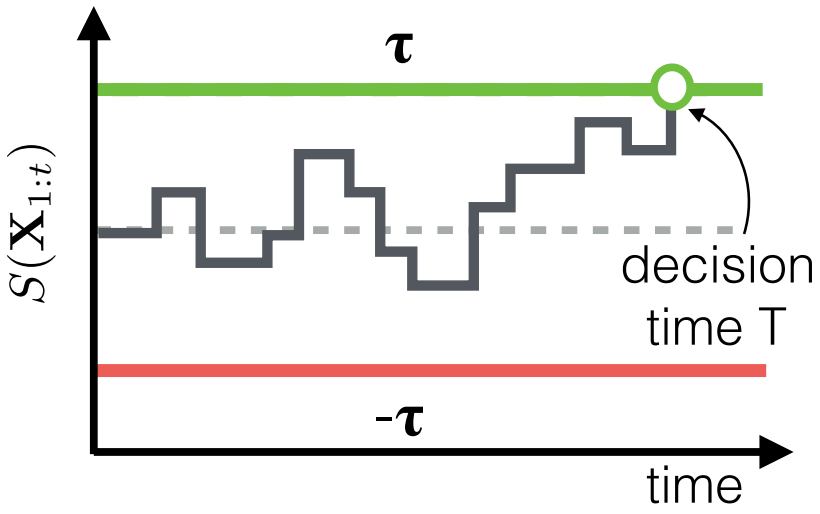

Figure 2.1: **The sequential probability ratio test (SPRT)**. SPRT **Eq. 2.3** computes the log class posterior ratio $S(\boldsymbol{X}_{1:t}) = \log \frac{P(C=1|\boldsymbol{X}_{1:t})}{P(C=0|\boldsymbol{X}_{1:t})}$ and compares to a pair of constant thresholds (assumed symmetrical here) for deciding whether to continue collecting observations and if not, which class prediction to make. The key in most applications is to compute $S(\boldsymbol{X}_{1:t})$.

> Here we assume that the two classes share the same prior probability of 0.5, hence the log posterior ratio $\log P(C = 1|\boldsymbol{X}_{1:t})/P(C = 0|\boldsymbol{X}_{1:t})$ is identical to the log likelihood ratio $\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)}$. If the prior probability is not uniform, one can obtain the log posterior ratio by adding the log prior ratio $\log \frac{P(C=1)}{P(C=0)}$, a simple application of Bayes' rule. Thus for simplicity, it is sufficient to be concerned with computing the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ only.

The thresholds τ and $-\tau$ are symmetrical as the class distributions and costs of errors are symmetrical. The threshold τ controls the maximum tolerable error rates. For example, if $\tau = 2$, i.e. predicting $C = 1$ when the object is $> 10^2$ times more likely to be in class 1 than in class 0, then the maximum error rate for misclassifying class $C = 1$ is 1%. Similarly If $\tau = 3$ then class 0 will be $< 10^3$ times more likely than class 1 when $C = 0$ is predicted, and the error rate for misclassifying $C = 0$ is at most 0.1%. τ is judiciously chosen by the observer to minimize the Bayes risk in **Eq. 2.1**, and hence is a function of the cost of error η.

To conclude, SPRT [1] essentially compares the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ between the two classes to a pair of thresholds τ and $-\tau$ that are constant over time. This simple algorithm enjoys optimality guarantees for a variety of classification tasks,

as we discuss below.

2.5 Optimality guarantees of SPRT

Simple hypothesis testing: strict optimality

SPRT is renowned for its optimality in "simple binary sequential testing" problems [1]. In these problems, the visible object belongs to one of two classes ($K = 2$), and given the class Y, the observations over time are independent and identically distributed (i.i.d.), i.e. $P(\boldsymbol{X}_{1:t}|C) = \prod_{t'=1}^{t} P(\boldsymbol{X}_{t'}|C)$. In this case Wald [1] proved that SPRT minimizes Bayes risk, i.e. any other sequential testing algorithm will either require longer reaction time or incur more error.

Composite hypothesis testing: asymptotic optimality

For more complex problems, SPRT has not been proven strictly optimal, but it often ensures "asymptotic" optimality, namely that its Bayes risk will be closer to optimal as error becomes more important (i.e. as $\eta \to \infty$). One such complex problem is binary composite hypothesis testing, where the object categories contain subclasses, and observations are i.i.d. given the subclasses, not the category C. In the lioness' problem, both categories ("weak" or "strong") are composite, e.g. a buffalo may be weak due to young/old age or past injuries, and the animal's appearance depends on these fine-grained subclasses. Composite hypothesis testing has been studied by many [7], [8] and shown to be asymptotically optimal: Lai [9] proves asymptotic optimality for a frequentist counterpart of the SPRT, and Darkhovsky [10] proves strict optimality in the minimax Bayesian setup. The other class of complex sequential testing problems is multi-hypothesis testing ($K \geq 2$, [11]–[13]), where SPRT has been shown to be asymptotic optimal[14].

How close to optimal is SPRT in non-asymptotic scenarios, i.e. (for finite η)? Strict optimality for SPRT in complex problems has not been obtained. Numerical simulations are therefore used to assess the performance of SPRTs on a problem specific basis (e.g. [8]). In **Ch. 6**, we provide optimality analysis of SPRT for the visual search problem (to be formally discussed in **Ch. 3**), and show that SPRT is near-optimal for most common settings.

2.6 Chapter summary

Our theoretical framework of choice is the sequential probability ratio test (SPRT). SPRT relies on thresholding a one-dimensional signal (the log posterior ratio) to determine the length of evidence accumulation and the final decision. SPRT achieves

impressive optimality guarantees for hypothesis testing problems where the hypotheses are (1) fully specified probabilistically and (2) static over time. In future chapters we will apply SPRT to vision problems with quantized inputs.

References

[1] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[2] E. R. Kandel, J. H. Schwartz, T. M. Jessell, *et al.*, *Principles of Neural Science*. McGraw-Hill New York, 2000, vol. 4.

[3] M. Meister and M. J. Berry, "The neural code of the retina," *Neuron*, vol. 22, no. 3, pp. 435–450, 1999.

[4] R. Vanrullen and S. J. Thorpe, "The time course of visual processing: From early perception to decision-making," *Journal of Cognitive Neuroscience*, vol. 13, no. 4, pp. 454–461, 2001.

[5] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

[6] J. R. Busemeyer and A. Rapoport, "Psychological models of deferred decision making," *Journal of Mathematical Psychology*, vol. 32, no. 2, pp. 91–134, 1988.

[7] T. Lai and D. Siegmund, "A nonlinear renewal theory with applications to sequential analysis i," *The Annals of Statistics*, pp. 946–954, 1977.

[8] G. Lorden, "Nearly-optimal sequential tests for finitely many parameter values," *The Annals of Statistics*, vol. 5, no. 1, pp. 1–21, 1977.

[9] T.-L. Lai, "Asymptotic optimality of generalized sequential likelihood ratio test in some classical sequential testing problems," *Sequential Analysis*, vol. 21, no. 4, pp. 219–247, 2002.

[10] B. Darkhovsky, "Optimal sequential tests for testing two composite and multiple simple hypotheses," *Sequential Analysis*, vol. 30, no. 4, pp. 479–496, 2011.

[11] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, 1994.

[12] A. G. Tartakovskii, "Sequential testing of many simple hypotheses with independent observations," *Problemy Peredachi Informatsii*, vol. 24, no. 4, pp. 53–66, 1988.

[13] G. Golubev and R. Khas' minskii, "Sequential testing for several signals in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 28, no. 3, pp. 573–584, 1984.

[14] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests. ii. accurate asymptotic expansions for the expected sample size," *Information Theory, IEEE Transactions on*, vol. 46, no. 4, pp. 1366–1383, 2000.

Chapter 3

VISUAL SEARCH

Sequential Reasoning with a Time-Homogeneous Probabilistic Model

We present a psychophysics study of visual search, which is concerned with explaining and assessing the optimality of human speed versus accuracy tradeoff (SAT). The advantage of psychophysics is that the experimenters, not nature, design the task, and therefore the probabilistic structure of the task is known. This project therefore showcases the power of our theoretical framework, the sequential probability ratio test (SPRT), when its assumptions are met (see **Ch. 2**), i.e. when the tasks can be fully specified probabilistically in a static environment.

3.1 The psychophysics of visual search

Visual search is the problem of looking for a target object amongst clutter or distractors. It is a common task for our everyday life (looking for keys on a desk, friends in a crowd or signs on a map) and a vital function for animals in the wild (searching for food, mate, threats). Visual search is difficult and error-prone: the sensory signal is often noisy; the relevant objects, and their appearance may not be entirely known in advance, are often embedded in irrelevant clutter, whose appearance and complexity may also be unknown. Thus to reduce detection errors the visual system must account for the noise structure of the sensors and the uncertainty of the environment. In addition, time is of the essence: the ability to detect quickly objects of interest is an evolutionary advantage. Speed comes at the cost of making more errors. Thus, it is critical that each piece of sensory information is used efficiently to produce a decision in the shortest amount of time while maintaining the probability of errors within an acceptable limit.

There are two crucial quantities in visual search: the **response time** (RT, how long after an observer is exposed to a scene before it generates a response) and the **error rate** (ER). The error rate includes the **false positive rate** (FPR), which is the fraction of times when the observer claims to have found a target even though the scene does not contain any, and the **false negative rate** (FNR), which is the fraction of times when the observer claims no target when there is one. We are interested in how these quantities are affected by the structure of the search task.

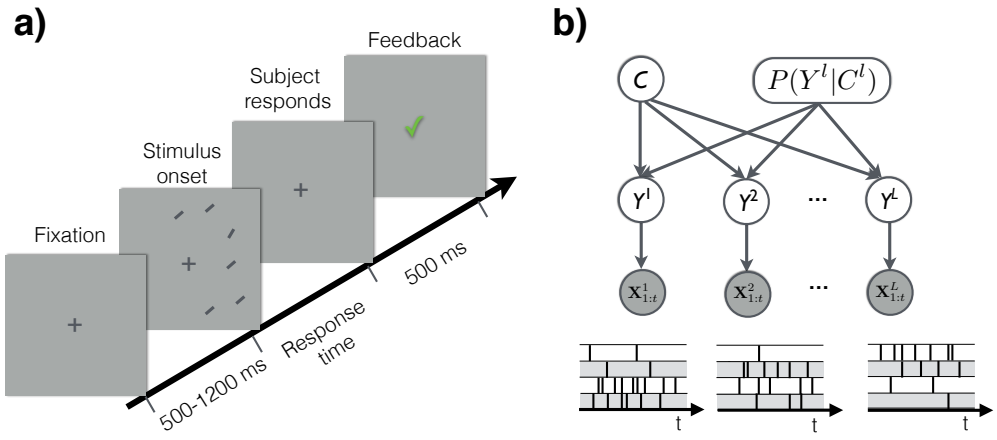

Figure 3.1: **Visual search setup** (**a**) Each trial starts with a fixation screen. Next, the "stimulus" is displayed. The stimulus is an image containing M oriented bars that are positioned in M out of L possible display locations ($M = 6$, $L = 12$ in this example). One of the bars may be the target. The stimulus disappears as soon as the subject responds by pressing one of two keys, to indicate whether a target was detected or not. Feedback on whether the response was correct is then presented on the screen, which concludes the trial. The subjects were instructed to maintain center-fixation at all times and respond as quickly and as accurately as possible. (**b**) A generative model of the stimulus. The stimulus class C and a prior distribution on the stimulus orientation $P(Y^l|C^l)$ decide, for each display location l, the orientation Y^l (may be blank). The orientation Y^l determines in turn the observations $\boldsymbol{X}^l_{1:t}$, which are firing patterns from a hypercolumn of V1 orientation-selective neurons at location l over the time window $[0, t\Delta]$ (The firing patterns of four neurons are shown at each location).

Psychologists have characterized human visual search performance [1]–[11] in relation to properties of the search environment such as the distinctiveness of the target against the background clutter [2], [3], the complexity of the image [4], [5] and the likelihood that an object of interest may be present [7], [9]. However, it is unknown what the optimal RT versus ER tradeoff should be in a given environment. It is also unknown whether human visual search performance is optimal.

Models of visual search fall into two categories. Stochastic accumulators were introduced to model discrimination [12]–[17] and visual search [18], [19]. The decision signal is either obtained from electrophysiological recordings from decision-implicated areas, e.g. frontal eye field [19]–[21] and lateral intraparietal area [22], [23]), or the result of an educated guess to fit the phenomenology [24], [25]. Stochastic accumulator models are appealing because of their conceptual simplicity and because they fit behavioral data well. However, these models do not attempt to

explain search performance in terms of the underlying primary signals and neural computations.

Ideal observer models have been developed to study which computations and mechanisms may be optimal for visual discrimination [24], [26] and visual search under fixed time presentations [27]–[31] using signal detection theory [32]. This line of work leads us to the question of whether it is possible to derive the optimal decision strategy for visual search that may predict simultaneously both RT and ER.

3.2 Contributions

We take the Bayesian point of view: we model a system that through experience (or through evolution) is familiar with the statistics of the scene. The input to our system is an array of idealized cortical hypercolumns that, in response to a visual stimulus, produce firing patterns that are Poisson and conditionally independent. After this assumption is made the model that characterizes the optimal ER vs RT tradeoff is derived with no additional assumptions and no additional free parameters.

Our main contributions are:

1. We propose a principled and parsimoneous model for studying **the optimal SAT of visual search**.
2. Our model can predict the observer's performance in **novel tasks** once some intrinsic properties of the input hypercolumn have been estimated.
3. We are interested in understanding whether such observer might be plausibly implemented by **neural mechanisms** such as a network of spiking neurons.
4. We assess the **optimality of humans** at visual search SAT. We collected psychophysics data and compare human performance with the optimal model and its spiking implementation.

3.3 Problem setup

The general set-up of a visual search task is as shown in **Fig. 3.1a**. An observer sits down in front of a computer monitor. The monitor displays a series of images that consists of distractors and sometimes targets. The goal of the observer is to decide whether a target object is present in a cluttered image as quickly and accurately as possible while maintaining fixation at the center of the image. The decision is binary, and the two categories of stimuli are: target-present ($C = 1$) and target-absent ($C = 0$), as shown in **Fig. 3.2a**. When the target is present, its location is not

known in advance; it may be one of L locations in the image. The observer only reports whether the target appears, but not where. For now, we limit the number of targets to be at most one.

In our experiments the target and distractor objects appear at M locations ($M \leq L$) in each image where M reflects the complexity of the image and is known as the **set-size**. The objects are simplified to be oriented bars, and the only feature by which the target and distractor differ is orientation. Target distinctiveness is controlled by the difference in orientation between target and distractors, the **orientation contrast** $\Delta\theta$. Prior to image presentation, the set of possible orientations for the target and the distractor is known, whereas the set-size and orientation contrast may be unknown, and may change from one image to the next (see **Fig. 3.2c-d** for examples).

> In this design we strive to have the simplest experiment that captures all the relevant variables, namely the dependent variables RT and ERs, as well as the independent variables the set-size M and the orientation contrast $\Delta\theta$. To do so we first simplify the appearance of the stimuli so that we can focus on modeling search strategies instead of building classifiers. Second, we eliminate eye-movements by forcing fixation at the center of the image at all times because saccade planning is a rich phenomenon on its own that many are struggling to explain. Third, we have randomized the placement of the targets and the distractors (details in **Sec. 3.5**), duration between trials, and stimulus orientation etc. to eliminate potential biases.

The visual search literature records a rich set of phenomena regarding the RT and ERs of human observers. We list three in **Fig. 3.3**. An intuitive phenomenon is the "set-size effect". As the amount of clutter increases in the display, the subject tends to take longer to respond. The slope of RT with respect to the set-size M depends on the distinctiveness between the target and the distractor $\Delta\theta$. The smaller $\Delta\theta$ is, the more difficult the task becomes and the larger the slope. A less intuitive phenomenon is the "search asymmetry effect" that the slope for target-absent is roughly twice the slope for target-present (many other dependent variables display the set-size effect and search asymmetry, the interested reader is referred to [4]). Lastly, the RT distributions is heavy-tailed: the log RTs roughly follow a Gaussian distribution. The list of phenomena goes on.

Existing visual search models [18], [19], [27], [28] describe a subset of the phenomena fairly well, but most fall short in accounting for phenomena across different

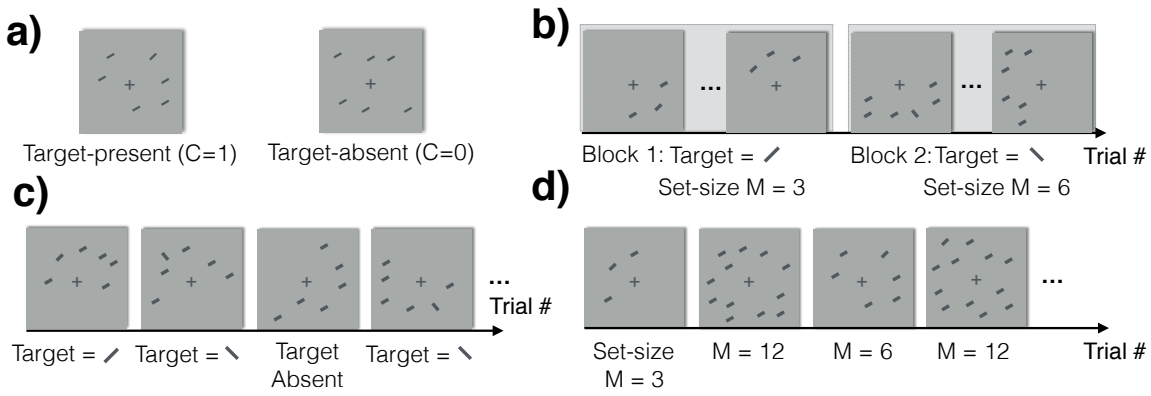

Figure 3.2: **Common visual search settings.** (**a**) The two stimulus categories: target-present and target-absent. (**b**) "Blocked": the orientation contrast and the set-size remain constant within a block of trials (outlined in gray boxes) and vary only between blocks. (**c**) "Mix contrast": the target orientation varies independently from trial to trial while the distractor orientation and the set-size are held constant. (**d**) "Mix set-size": the set-size is randomized between trials while the target and distractor orientations are fixed.

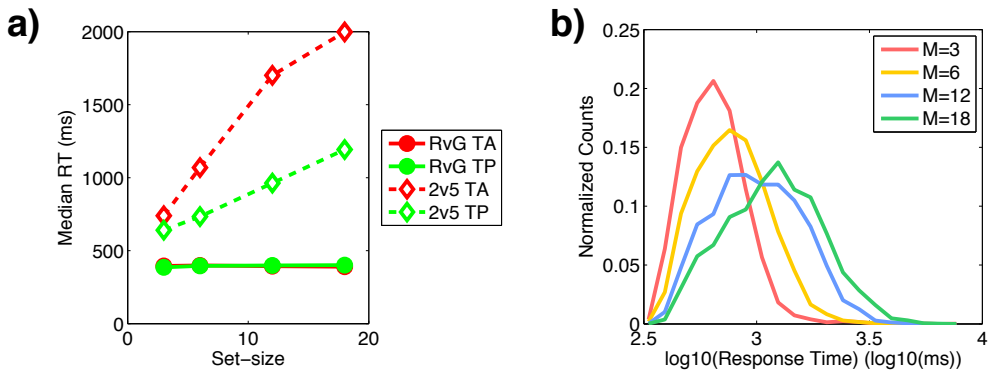

Figure 3.3: **Selected list of visual search phenomena** (**a**) The "set-size" effect. Median RT increases linearly with set-size. The slope depends on the trial type (target-absent trials have roughly twice the slope) and task difficulty. The two tasks are searching for a red bar among green bars (easy) and searching for a "2" among "5"s (hard). (**b**) RT histograms for different set-sizes ($\{3, 6, 12, 18\}$), plotted in log domain based 10.

search environments. Describing all phenomena in one model is a challenging task. The model needs to be flexible enough to accommodate changes of the environment, e.g. different set-sizes, or different probability distributions on the set-sizes, etc. In addition, the model needs to be efficient enough so that it can be easily transferred from one environment to the next. Furthermore, there are countless unintended events, such as the subject blinking, getting fatigued or being distracted, that could

pollute the behavioral data.

Therefore, instead of **describing** human behaviors in a variety of visual search problems, we seek to study the **optimal** behavior on a per-situation basis. The optimal behavior can be used as a gold standard to measure human performance. Given input observations and prior knowledge about the task, we are interested in the best achievable ER versus RT tradeoff measured in Bayes risk (**Eq. 2.1**).

3.4 Asymptotically optimal search model

Quantized sensory input

The first step towards studying optimal SAT is to identify the input to the problem. We consider sensory input from the early stages of the visual system (retina, lateral geniculate nucleus (LGN) and primary visual cortex), where raw images are processed and converted into a stream of quantized events, aka **action potentials**. The anatomy, as well as the physiology, of these stages are well characterized [33]. These mechanisms compute local properties of the image, such as color contrast, orientation, spatial frequency, stereoscopic disparity and motion flow[34], and communicate these properties to downstream neurons for further processing. The communication takes on the forms of sequences of action potentials / spikes from orientation-selective neurons in V1 [33].

The firing patterns of the neurons are modeled with an homogeneous Poisson process [35]. This means that each neuron fires at a fixed rate of λ spikes / second given the input image, and the timings of the spikes are independent of each other. More specifically, the number n of events (i.e. action potentials) that will be observed during one second is distributed as

$$P(n|\lambda) = \lambda^n e^{-\lambda}/n!.$$

The firing patterns $\boldsymbol{X}_{1:t}$ are produced over the time interval $[0, t\Delta]$ by a population of n_H neurons, also known as a **hypercolumn**, from each of the L display locations. We model each neuron using the Linear Nonlinear Poisson (LNP) model [36], [37], which is commonly used to model neural responses. Each neuron has a localized spatial receptive field and is tuned to local image properties [33], which in our case is the local stimulus orientation; the preferred orientations of neurons within a hypercolumn are distributed uniformly in $[0°, 180°)$. λ^i_θ, the expected firing rate of the i-th neuron, is a function of the neuron's preferred orientation θ_i and the stimulus orientation $\theta \in [0°, 180°)$:

$$\lambda_\theta^i = (\lambda_{max} - \lambda_{min}) \exp\left(-\frac{||\theta - \theta_i||^2}{\sigma_Y^2}\right) + \lambda_{min}, \tag{3.1}$$

(in spikes per second, or Hz) where λ_{min} and λ_{max} are a neuron's minimum and maximum firing rates, $||\theta - \theta_i||$ denotes the minimum angular distance between θ and θ_i, and $\sigma_Y \in (0°, 180°)$ is the half tuning width. **Fig. 3.4a** shows the tuning functions of a hypercolumn of eight neurons, **Fig. 3.4b** shows the spatial organization of the hypercolumns, and **Fig. 3.4c-d** shows the sample spike trains from two locations with different local stimulus orientations.

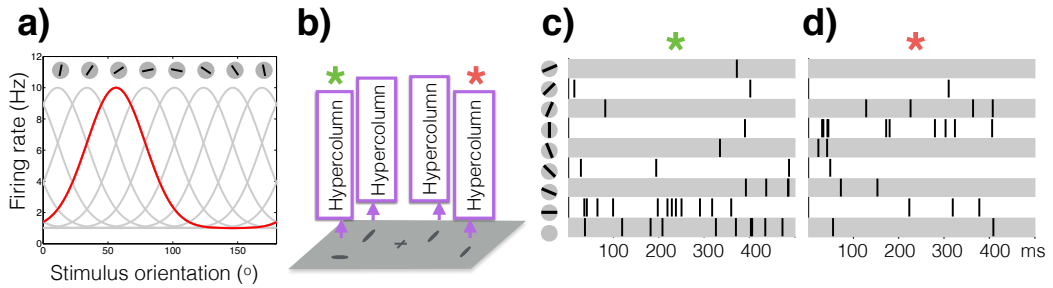

Figure 3.4: **V1 Hypercolumns** (**a**) Orientation tuning curves λ_θ^i (**Eq. 3.1**) of a hypercolumn consisting of $n_H = 8$ neurons with half tuning width $\sigma_Y = 22°$, minimum firing rate $\lambda_{min} = 1Hz$ and maximum firing rate $\lambda_{max} = 10Hz$. (**b**) V1 hypercolumns tessellate the input space, one for each visual location where an object (oriented bar) may appear. (**c-d**) Spike trains $\boldsymbol{X}_{1:t}^l$ at the target location (marked with green star in (b)) and a distractor location (red star).

Why do we select the response of V1 hypercolumn neurons to be our input? Indeed there are multiple alternatives: the raw image, the response of the retina or LGN, and high-level signals that directly encode information regarding target presence. Our choice is based on flexibility and efficiency. Since the search problems considered here all involve a simple scenario of oriented bars placed certain distances apart, it would be redundant to model the neuronal hardware that gives rise to orientation-selectivity at this stage. Therefore, our level of abstraction should start at least from V1. On the other hand, although most visual search models assume high-level input signals [18], [19], [27], [28], they are not concerned with behaviors across multiple visual search tasks. As we see later, we will interpret the input from V1 neurons depending on the probabilistic structure of the task, which is key for SPRT to generalize across tasks.

> Why do we use LNP to model the V1 spike trains? While Gaussian firing rate models [28] have also been used in the past, the Poisson model represents more faithfully the spiking nature of neurons [35], [38], [39]. Second, the LNP model is simple and parsimonious: it is well studied in the literature [40], and its limitations are increasingly well understood [40]. Lastly, we do not use electrophysiological recordings from V1 neurons [39] because large-scale recordings from the entire V1 are not currently possible. Nonetheless, it may be possible to bootstrap from a well-represented population of V1 neurons.

Sequential probability ratio test for visual search

Since the problem is binary, SPRT (**Eq. 2.3**) applies directly to the quantized spike-train input $\boldsymbol{X}_{1:t}$ of V1 hypercolumn neurons from all display locations over duration $[0, t\Delta]$:

$$S(\boldsymbol{X}_{1:t}) \triangleq \log \frac{P(C=1|\boldsymbol{X}_{1:t})}{P(C=0|\boldsymbol{X}_{1:t})} \begin{cases} \geq \tau_1 & \text{Declare target present} \\ \leq \tau_0 & \text{Declare target absent} \\ \text{otherwise} & \text{Postpone decision,} \end{cases} \tag{3.2}$$

where $S(\boldsymbol{X}_{1:t})$ is the log likelihood ratio of target-present ($C = 1$) vs. target-absent ($C = 0$) probabilities with respect to the observations $\boldsymbol{X}_{1:t}$. τ_1 and τ_0 together control the maximum false positive and false negative rates. The key to applying SPRT is to compute $S(\boldsymbol{X}_{1:t})$, which may be systematically constructed from the visual input according to the graphical model in **Fig. 3.1b**, and can account for a wide variety of visual search tasks.

> We derive a general model that is capable of handling unknown set-sizes and orientation contrasts. To build up the concept, we start by reviewing models for simpler tasks including visual discrimination and visual search with known set-sizes and orientation contrasts, both of which have already been explored in the literature [29], [41], [42]. Readers only interested in this general model are encouraged to skip these models. **Table 3.1** provides a roadmap for the models.

Chapter-specific notations

Let $\boldsymbol{X}_t^l$ denote the activity of the neurons at location l during the time interval $[0, t\Delta]$ in response to a stimulus presented at time 0. $\boldsymbol{X}_{1:t} = \{\boldsymbol{X}_t^l\}_{l=1}^L$ is the ensemble responses of all neurons from all locations. Let $\mathcal{L}_\theta(\boldsymbol{X}_{1:t}^l) \triangleq \log P(\boldsymbol{X}_{1:t}^l|Y^l = \theta)$ denote the log likelihood of the spike train data $\boldsymbol{X}_{1:t}^l$ when the object orientation

Task	L	M	$\Delta\theta$	CCD	Expression
Homogeneous discrimination	1	$M = 1$	known	known	**Eq. 3.3**
Heterogeneous discrimination	1	$M = 1$	unknown	known	**Eq. 3.5**
Homogeneous search	> 1	$M = L$	known	known	**Eq. 3.7**
I.i.d-distractor hetero-search	> 1	$M = L$	unknown	known	**Eq. 3.8**
Heterogeneous search	> 1	unknown	unknown	unknown	**Eq. 3.10**

Table 3.1: **List of visual discrimination and visual search tasks**. Our contribution is developing models for tasks colored in blue. In addition, our general model accounts for the heterogeneous search task, which subsumes all other tasks on the list. L is the number of total display locations. M is the number of display items. θ_T and θ_D are the target and distractor orientations, respectively. We use "known" and "unknown" to refer to whether a quantity is known at stimulus onset. In many tasks, θ_T and θ_D are unknown, but sampled according to a distribution. The distribution ϕ of the distractor orientation is called a conditional distractor distribution (CDD , see the i.i.d-heterogeneous search section), where $\phi_\theta = P(Y^l = \theta | C^l = 0)$ for any location l. $S(\boldsymbol{X}_{1:t}) = \log P(C = 1|\boldsymbol{X}_{1:t})/P(C = 0|\boldsymbol{X}_{1:t})$ is the class log posterior ratio that SPRT computes.

Y^l at location l is θ (degrees). When there is only one location (as in visual discrimination as below), the location superscript is omitted. The target orientation and the distractor orientation are denoted respectively by θ_T and θ_D. In many cases, the target orientation is not unique, but sampled from a set $\Theta_T = \{\theta_1, \theta_2, \ldots\}$ of many possible values. Simiarly Θ_D is the domain for the distractor orientation. $n_T = |\Theta_T|$ and $n_D = |\Theta_D|$ are the number of candidate target and distractor orientations, respectively.

Homogeneous visual discrimination

First consider the case where either the target or the distractor can appear at only one display location ($L = M = 1$), and the target and distractor have distinct and unique orientations, θ_T and θ_D, respectively. The visual system needs to determine whether the target or the distractor is present in the test image. The log likelihood ratio in this case is well known [41] (re-derived in the Appendix (**Eq. A.3**)):

$$\text{(Homogeneous Discrimination) } S(\boldsymbol{X}_{1:t}) = \mathcal{L}_{\theta_T}(\boldsymbol{X}_{1:t}) - \mathcal{L}_{\theta_D}(\boldsymbol{X}_{1:t}), \tag{3.3}$$

which, as first pointed out by [43], may be computed by a diffuse-to-bound mechanism [12]. $S(\boldsymbol{X}_{1:t})$ is a 'diffusion', i.e. it can be updated additively (see **Eq. 3.13**):

$$S(\boldsymbol{X}_{1:t}) = S(\boldsymbol{X}_{1:t-1}) + \left(\mathcal{L}_{\theta_T}(X_t) - \mathcal{L}_{\theta_D}(X_t)\right), \tag{3.4}$$

and a decision is taken whenever the diffusion hits one of two boundaries, hence the name "diffuse-to-bound". In addition, as shown by [41], SPRT is optimal in minimizing the Bayes risk in **Eq. 2.1**.

Heterogeneous visual discrimination

In a more general setting, both the target and the distractor could take one of multiple orientations. We call heterogeneous visual discrimination the case where the target and distractors could take on one of multiple orientations, i.e. $n_T > 1$ and/or $n_D > 1$. The log likelihood ratio is [29] (re-derived in Appendix (**Eq. A.4**)):

> How much does the form of $S(\boldsymbol{X}_{1:t})$ depend on the observations $\boldsymbol{X}_{1:t}$ being Poisson? Only $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ makes use of the Poisson likelihood, the derivation of $S(\boldsymbol{X}_{1:t})$ based on $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ simply follows Bayesian inference and is therefore independent of the form of the observation likelihood.

$$\begin{aligned}\text{(Heterogeneous Discrimination) } S(\boldsymbol{X}_{1:t}) = &\underset{\theta\in\Theta_T}{\mathcal{S}\text{max}}\left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) - \log(n_T)\right) \\ &- \underset{\theta\in\Theta_D}{\mathcal{S}\text{max}}\left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) - \log(n_D)\right),\end{aligned} \tag{3.5}$$

where $\mathcal{S}\text{max}\,(\cdot)$ is the "softmax" function. For a vector $\mathbf{v}$ and a set of indices $\mathcal{I}$:

$$\underset{i\in\mathcal{I}}{\mathcal{S}\text{max}}\,(\mathbf{v}) \triangleq \log\sum_{i\in\mathcal{I}}\exp(v_i). \tag{3.6}$$

Softmax can be thought of as the marginalization operation in log probability space: it computes the log probability of a set of mutually-exclusive events from the log probabilities of the individual events. For example, for two mutually-exclusive events, A_1 and A_2, we have $P(A_1 \bigcup A_2) = P(A_1) + P(A_2)$, then $\log P(A_1 \bigcup A_2) = \underset{i=1,2}{\mathcal{S}\text{max}}\,(\log P(A_i))$. Since the different target orientations are mutually-exclusive, their log likelihoods should be combined using the softmax function to compute the log likelihood for the target. The same argument applies to the distractor.

It is important to note that the log likelihood ratio for heterogenous discrimination is **not a diffusion**, as **Eq. 3.5** does not admit an additive update formulation as in **Eq. 3.4**. Rather, it combines diffusions in a non-linear fashion (via a softmax). Diffuse-to-bound [12] does **not** give the optimal decision mechanism here, nor in any of the settings we will discuss later. Moreover, while a diffusion model may require additional parameters specifying how the statistics of the diffusions relate to the task parameters (set-size in this case) [24], [25], the construction of SPRT is *parameter-free*. Later in **Fig. 3.10c-f** we will see that SPRT can generalize to novel experimental settings. The generalizability is non-trivial for diffusion models.

Homogenous search

Now that we have analyzed the case of discrimination (one item visible at any time) we will explore the case of search (multiple items present simultaneously, one of which may be the target). Consider the case where all the L display locations are occupied by either a target or a distractor (i.e. $L = M > 1$) and the display either contains one target or none. The target orientation θ_T and the distractor orientation θ_D are again unique and known, i.e. $n_T = n_D = 1$. The log likelihood ratio of target-present vs target-absent is given by [42] (re-derived in Appendix **Eq. A.5**):

$$\text{(Homogeneous Search) } S(\boldsymbol{X}_{1:t}) = \underset{l=1,\ldots,L}{\mathcal{S}\max} \left(S(\boldsymbol{X}^l_{1:t}) - \log(L) \right), \tag{3.7}$$

where $S(\boldsymbol{X}^l_{1:t}) = \mathcal{L}_{\theta_T}(\boldsymbol{X}^l_{1:t}) - \mathcal{L}_{\theta_D}(\boldsymbol{X}^l_{1:t})$ is the log likelihood ratio for homogenous discrimination at location l (see **Eq. 3.3**). $S(\boldsymbol{X}_{1:t})$ combines the local log likelihood ratio $S(\boldsymbol{X}^l_{1:t})$ from all locations using a softmax because the target can only appear at one of L disjoint locations.

I.i.d.-distractor heterogeneous search

Now we describe our general model of visual search. We start with the simple case where the set-size is known ($M = L > 1$) but the orientation contrast is not ($n_T > 1$, and/or $n_D > 1$). In addition, we assume target and distractor orientations are sampled **i.i.d.** in space according to some distribution. We refer to this as the i.i.d.-distractor heterogeneous search.

We call a "conditional distractor distribution" (CDD) the distribution of orientation Y^l at any non-target location l, i.e. $P(Y^l|C^l = 0)$. We denote CDD with ϕ where $\phi_\theta \triangleq P(Y^l = \theta|C^l = 0)$. Thus ϕ is a n_D-dimensional probability vector. i.e. each

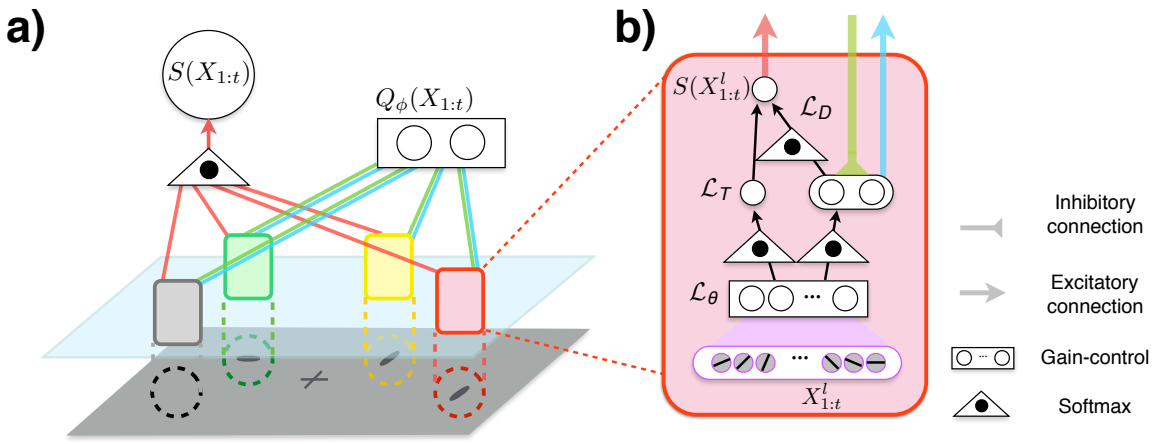

Figure 3.5: **SPRT for heterogeneous visual search.** (**a**) SPRT for heterogeneous visual search is implemented by a five-layer network. It has two global circuits, one computes the global log likelihood ratio $S(\boldsymbol{X}_{1:t})$ (**Eq. 3.10**) from local circuits that compute log likelihood ratios $\{S(\boldsymbol{X}^l_{1:t})\}_l$ (**Eq. 3.11**), and the other estimates scene complexity $Q_\phi(\boldsymbol{X}_{1:t})$ (**Eq. A.9**) via gain-control. $Q_\phi(\boldsymbol{X}_{1:t})$ feeds back to the local circuit at each location. (**b**) The local circuit that computes the log likelihood ratio $S(\boldsymbol{X}^l_{1:t})$. Spike trains $\boldsymbol{X}_{1:t}$ from V1/V2 orientation-selective neurons are converted to log likelihood for task-relevant orientations $\mathcal{L}_\theta$ (**Eq. 3.13**). The log likelihoods of the distractor $\mathcal{L}_D$ (second line of **Eq. 3.9**) under every putative CDD are compiled together, sent (blue outgoing arrow) to the global circuit, and inhibited (green incoming arrow) by the CDD estimate Q_ϕ (details in **Eq. A.9**).

element of ϕ is non-negative, and all elements sum to one. We introduce CDD here because it is a key element in the general model of visual search, as will become clear later. In contrast, the conditional target distribution $P(Y^l = \theta | C^l = 1)$ is not as vital and is assumed uniform for notation clarity (see Appendix **Eq. A.11** for cases with general target distributions and different CDDs over locations, and see Appendix **Sec. A.1** for how to formulate common search problems such as those illustrated in **Fig. 3.2b-d** in the framework using CDDs.).

The log likelihood ratio may be computed as:

$$\text{(I.i.d.-Distractor Heterogeneous Search) } S(\boldsymbol{X}_{1:t}) = \underset{l=1\ldots L}{\mathcal{S}\text{max}} \left(S(\boldsymbol{X}^l_{1:t}) - \log(L) \right), \tag{3.8}$$

$$\text{where } S(\boldsymbol{X}^l_{1:t}) = \underset{\theta\in\Theta_T}{\mathcal{S}\text{max}} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) - \log(n_T) \right) - \underset{\theta\in\Theta_D}{\mathcal{S}\text{max}} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) + \log \phi_\theta \right). \tag{3.9}$$

The log likelihood ratio expressions (**Eq. 3.8 -3.9**) are obtained by nesting appropriately the models of homogeneous search and heterogeneous discrimination. At the highest level is the softmax over locations as in **Eq. 3.7**. At each location l, $S(\boldsymbol{X}^l_{1:t})$ is obtained as the difference between the log likelihood of the target with that of the distractor (**Eq. 3.9**), which is reminiscent of **Eq. 3.5**. Computing the target log likelihood requires marginalizing over the unknown target orientation with a softmax (again assuming uniform prior over possible target orientations in Θ_T). Similarly, the distractor log likelihood marginalizes over the distractor orientation according to the CDD.

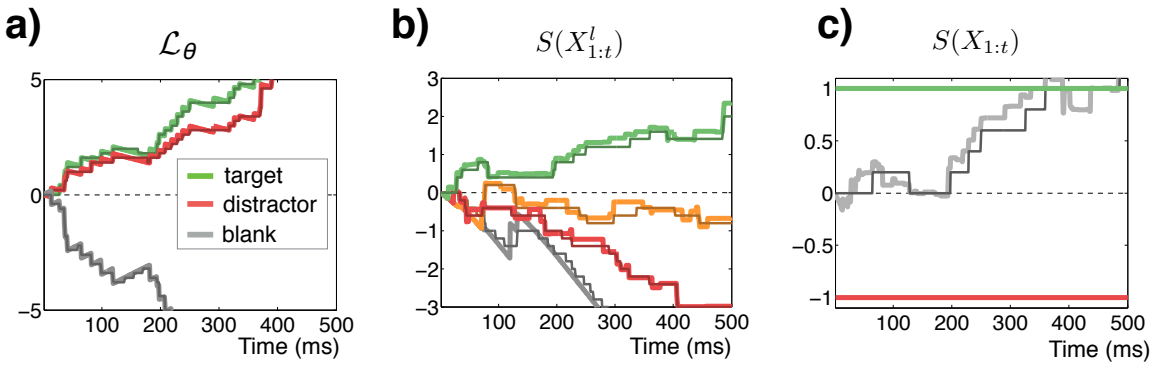

Figure 3.6: **An instantiation of the signals propagating through the network in Fig. 3.5a.** The orientation contrast is 45° and there are two possible set-sizes, 1 and 3. (**a**) The orientation log likelihoods $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ (**Eq. 3.13**) at the target location (green box in **Fig. 3.5a**). Lighter colors correspond to the analog signal and darker colors correspond to the spiking network approximation. (**b**) Local log likelihood ratios $S(\boldsymbol{X}^l_{1:t})$ (**Eq. 3.11**) for the four color-coded locations in **Fig. 3.5a**. (**c**) the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ (**Eq. 3.10**) computed using SPRT (black line) and the spiking implementation (gray line) reach the identical decision at similar response times (350ms).

Heterogeneous search

Finally, in the most ecologically relevant situations the complexity and target distinctiveness are not known in advance. In other words, all search parameters M, θ_T and θ_D are stochastic (n_T and/or $n_D > 1$). This scenario may be handled using the mechanisms for i.i.d. distractor heterogeneous search above as building blocks. For example, for a fixed set-size, each non-target location has a certain probability of being blank (as oppose to containing a distractor), which is captured by the CDD. When set-size changes, CDD will change correspondingly. Therefore, knowing the CDD effectively allows us to infer the set-size, and vice versa. Our strategy is to infer the CDD along with the class variables using Bayesian inference.

Let $P(\phi)$ be the prior distribution over the CDDs ϕ. Note that, technically, $P(\phi)$ is a "distribution over distributions". Computing the log likelihood ratio requires marginalizing out ϕ according to $P(\phi)$ and the observation $\boldsymbol{X}_{1:t}$. We assume that the observer has been exposed to this task for some time and has estimated $P(\phi)$. We also assume that the target distribution is independent of the CDD (and relax this assumption in the Appendix **Eq. A.14**). The log likelihood ratio is (see derivations in Appendix **Eq. A.14**):

$$\text{(General Model: Heterogeneous Search) } S(\boldsymbol{X}_{1:t}) = \underset{l=1\ldots L}{\mathcal{S}\max}\left(S(\boldsymbol{X}^l_{1:t}) - \log(L)\right), \tag{3.10}$$

$$\begin{aligned}\text{where } \; S(\boldsymbol{X}^l_{1:t}) = {} & \underset{\theta\in\Theta_T}{\mathcal{S}\max}\left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) - \log(n_T)\right) \\ & + \underset{\phi\in\Phi}{\mathcal{S}\max}\left(-\underset{\theta\in\Theta_D}{\mathcal{S}\max}\left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) + \log\phi_\theta\right) + Q_\phi(\boldsymbol{X}_{1:t})\right),\end{aligned} \tag{3.11}$$

where $Q_\phi(\boldsymbol{X}_{1:t}) \triangleq \log P(\phi|\boldsymbol{X}_{1:t})$ is the log posterior of the CDDs given the observations $\boldsymbol{X}_{1:t}$ (see below). The only difference between the equations **Eq. 3.10 -3.11** and those describing the i.i.d.-distractor heterogeneous search (**Eq. 3.8 -3.9**) is the second line of **Eq. 3.11**, where the CDD is marginalized out with respect to $Q_\phi(\boldsymbol{X}_{1:t})$. Since both the CDD ϕ and the distractor orientation Y^l must be marginalized, two softmaxes are necessary (the second line of **Eq. 3.11**). The equations do not explain how to compute $Q_\phi(\boldsymbol{X}_{1:t})$. It may be estimated simultaneously with the main computation by a scene complexity mechanism that is derived from first principles of Bayesian inference (see Appendix **Eq. A.9**). This mechanism extends across the visual field and may be interpreted as wide-field gain-control (see **Fig. 3.5a**).

A simpler alternative to inferring the CDD on a trial-by-trial basis is to ignore its variability completely by always using the same CDD obtained from the average complexity and target distinctiveness. More specifically, the approximated log likelihood ratio is:

$$\tilde{S}(\boldsymbol{X}_{1:t}) \approx \underset{l=1,\ldots,L}{\mathcal{S}\max} \left(\underset{\theta\in\Theta_T}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) \right) - \underset{\theta\in\Theta_D}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) + \log \bar{\phi}_\theta \right) \right) - \log(n_T L), \tag{3.12}$$

where $\bar{\phi}_\theta = \mathbb{E}(\phi_\theta)$ is the mean CDD for orientation θ with respect to the its prior distribution. This approach is suboptimal. Intuitively, if the visual scene switches randomly between being cluttered and sparse, then always treating the scene as if it had medium complexity would be either overly-optimistic or overly-pessimistic. Crucially, the predictions of this simple model are inconsistent with the behavior of human observers, as we shall see later in **Fig. 3.8**.

3.5 Model prediction and human psychophysics

Now that we have seen how to implement SPRT given a visual search task, we show that it can predict existing phenomena in the literature and data collected by ourselves.

Qualitative fits

A first test of our model is to explore its qualitative predictions of RT and ER in classical visual search experiments (**Fig. 3.1a**).

In a first simulation experiment (**Sim. 1**), we used a "blocked" design (**Fig. 3.2b**), where the orientation of targets and distractors as well as the number of items do not change from image to image within an experimental block. Thus, the observer knows the value of these parameters from experience. Accordingly, we held these parameters constant in the model. We assume that the costs of error are constant, hence we hold the decision thresholds constant as well. What changes from trial to trial is the presence and the location of the target, and the timing of individual action potentials in the simulated hypercolumns. Since we do not model eye-fixations, we assume that the observer can see all the items equally (which corresponds to enforcing fixation at the center of the screen for human subjects).

The model makes three qualitative predictions: (a) The RT distribution predicted by the model is heavy-tailed: it is approximately log-normal in time (**Fig. 3.7b**).

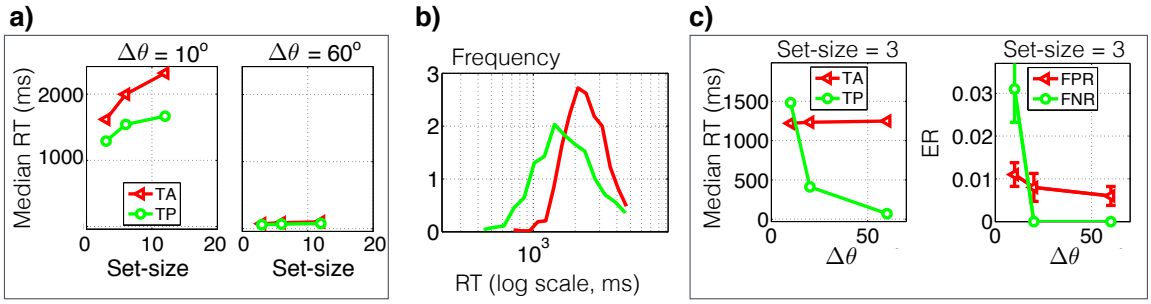

Figure 3.7: **Qualitative predictions of SPRT (Sim. 1-2).** (**a**) Set-size effect on median RT under the blocked design (**Sim. 1**). The ideal observer predicts a linear RT increase with respect to set-size when the orientation contrast $\Delta\theta$ is low (10°, left) and a constant RT when the orientation contrast is high (60°, right). The target-absent (TA) RT slope is roughly twice that of target-present (TP). (**b**) RT histogram under the blocked design with a 10° orientation contrast and a set-size of 12 items. RT distributions are approximately log-normal. (**c**) Median RT (upper) and ER (lower) for visual search with heterogenous target/distractor, mixed design (**Sim. 2**).

(b) The median RT increases linearly, as a function of M, with a large slope for hard tasks (small orientation contrast between target and distractor), and almost flat for easy tasks (large orientation contrast) (**Fig. 3.7a**). The median RT is longer for target-absent than for target-present, with roughly twice the slope (**Fig. 3.7a**). The three predictions are in agreement with classical observations in human subjects (**Fig. 3.3**) [1], [44].

In a second experiment (**Sim. 2**) we adopted a "mixed" design, where the distractors are known, but the orientation contrast is sampled from 10^0, 20^0 and 60^0, randomized from image to image (**Fig. 3.2c**). The subjects (and our model) do not know which orientation contrast is present before stimulus onset. The predictions of the model are shown in **Fig. 3.7c**. When the target is present both RT and ER are sensitive to the orientation contrast and will decrease as the orientation contrast increases, i.e. the model predicts that an observer will trade off errors in difficult trials (more errors) with errors in easy trials (fewer errors) to achieve an overall desired average performance, which is consistent with psychophysics data.

In **Sim. 3** we explored which one of two competing models best accounts for visual search when scene complexity is unknown in advance **Fig. 3.7d**). Recall that in discussing the heterogeneous search we proposed two models, one that estimates scene complexity (**Eq. 3.10**) and is optimal, and a simplified model (**Eq. A.13**) that is sub-optimal. The optimal model predicts that ERs are comparable for different set-

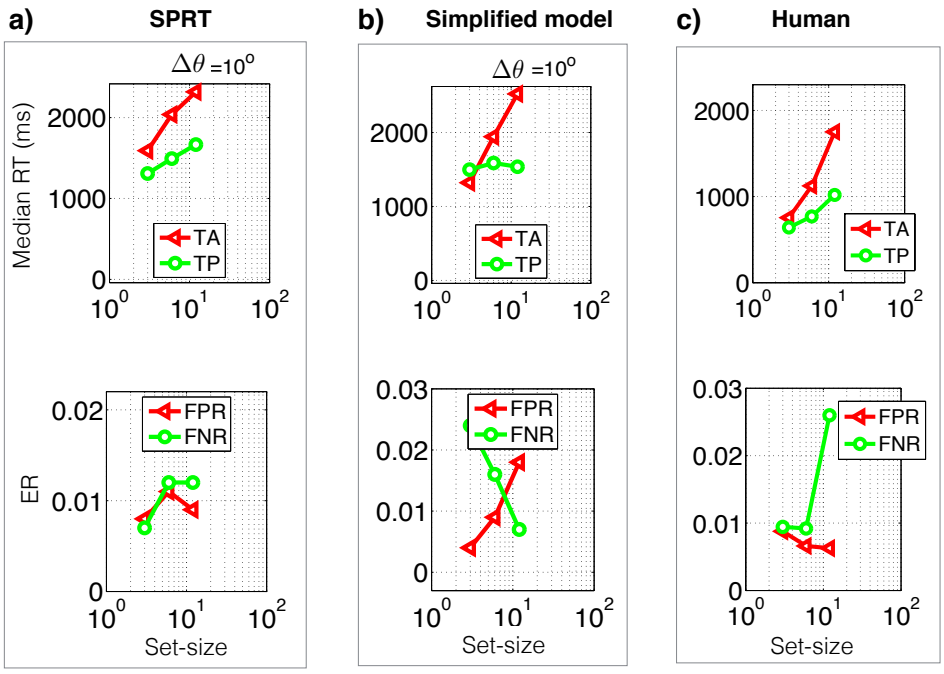

Figure 3.8: **Qualitative model predictions and psychophysics data on visual search with unknown set-size (Sim. 3).** Median RT (upper) and ER (lower): false-positive-rate (FPR) and false-negative-rate (FNR), of visual search with homogenous target/distractor and unknown set-sizes (**Sim. 3**) under two models: SPRT (**a**) that estimates the scene complexity parameter ϕ (essentially the probability of a blank at any non-target location) on a trial-by-trial basis (**Eq. 3.10**) using a wide-field gain-control mechanism (**Eq. A.9**); and a simplified observer (**b**) that uses average scene complexity $\hat{\phi}$ for all trials (**Eq. A.13**). Psychophysical measurements on human observers (Wolfe et al. [9], spatial configuration search in Fig. 2-3, reproduced here as (**c**)) are consistent with the optimal model (**a**). Simulation parameters are identical to those used in **Fig. 3.7.**

sizes while RTs show strong dependency on set-size when the orientation contrast is small (**Fig. 3.8a**). The simplified model, where scene complexity is assumed constant (**Eq. A.13**), predicts the opposite, i.e. that ER will depend strongly on set size, while RT will be almost constant when the target is present (**Fig. 3.8b**). Human psychophysics data ([9], reproduced in **Fig. 3.8c**) show a positive correlation between RT and set-size and little dependency of ER on set-size, which favor the optimal model and suggest that the **human visual system estimates scene complexity** while it carries out visual search.

Quantitative fits

In order to assess our model quantitatively, we compared its predictions with data harvested from human observers who were engaged in visual search (**Fig. 3.1a**). Three experiments were conducted to test both the model and humans under different conditions. The conditions are parameterized by the orientation contrast chosen from $\{20°, 30°, 45°\}$ and the set-size chosen from $\{3, 6, 12\}$. The blocked design was used in the first experiment (**Exp. 1**), where all $3 \times 3 = 9$ pairs of orientation contrast and set-size combinations were tested in blocks. The second experiment randomized orientation contrast from trial to trial while fixing the set-size at 12 (**Exp. 2**). The third randomized the set-size while holding the orientation contrast fixed at 30° (**Exp. 3**). The subjects were instructed to maintain eye-fixation at all times, and respond as quickly as possible and were rewarded based on accuracy.

We fit our model to explain the **full RT distributions and ERs** for each design separately. In order to minimize the number of free parameters, we held the number of hypercolumn neurons constant at $n_H = 16$, their minimum firing rate constant at $\lambda_{\min} = 1Hz$, and the half-width of their orientation tuning curves at 22° (full width at half height: 52°) [39]. Hence we were left with only three free parameters: the maximum firing rate of any orientation-selective neuron λ_{max} controls the signal-to-noise ratio of the hypercolumn; the upper and lower decision thresholds τ_0 and τ_1 control the frequency of false alarm and false reject errors. Once these parameters are given, all the other parameters of our model are analytically derived.

While our model takes care of the perceptual computational time, human response times also include a non-perceptual motor and neural conduction delay [44]. Therefore, we also use two additional free parameters per subject to account for the non-perceptual delay. We assume that the delay follows a log-normal distribution parameterized by its mean and variance.

In the blocked design experiment **Exp. 1**, the hypercolumn and the motor time parameters were fit jointly across all blocks (about 1620 trials); the decision thresholds were fit independently on each block (180 trials/block). In the mixed design experiments **Exp. 2-3**, all five parameters were fit jointly across all conditions for each subject because all conditions are mixed (440 trials/ condition). See **Fig. 3.9** for data and fits of a randomly selected individual, and **Fig. 3.10a-b** for all subjects in the blocked condition. In each experiment the model is able to fit the subjects' data well. The parameters that the model estimated (the maximum firing rate of the neurons λ_{max}, the decision thresholds τ_0, τ_1 are plausible [45]). Each subject

displays different ERs for different conditions (see **Fig. 3.11**), and thus the decision thresholds are indeed not constant.

> It may be possible to model the inter-condition variability of the thresholds as the result of the subjects minimizing a global risk function [25]. Therefore for each subject in the blocked design experiment **Exp. 1** we have tried fitting a common Bayes risk function (**Eq. 2.1**), parameterized by the two costs of errors, η_0 and η_1, across all blocks, and solving for the optimal thresholds for each block independently. This assumption reduces the number of free parameters for the blocked condition from 21 (2 thresholds $\times$ 9 conditions + 1 SNR + 2 motor parameters) to 5 (2 costs of errors + 1 SNR + 2 motor parameters), but at the cost of marked reduction in the quality of fits for some of the subjects. Therefore as far as our model is concerned, there was some block-to-block variability of the error costs.

Finally, we test our model's generalization ability. We used the signal-to-noise ratio parameter (the maximum firing rate λ_{max}) and the two non-decision delay parameters estimated from the blocked experiment (**Exp. 1**) to predict the mixed experiments (**Exp. 2-3**). Thus for each mixed experiment only two parameters, namely the decision thresholds τ_0 and τ_1, were fit. Despite the parsimony in parameterization, the model shows good cross-experiment fits (see **Fig. 3.10c-f**), suggesting that the parameters of the model refer to real characteristics of the subject.

In conclusion, SPRT both prescribes the optimal behavior given task structure and predicts human visual search behavior. SPRT has a compact parameterization: on average, three parameters are needed to predict each experimental condition and many parameters (the signal-to-noise ratio of the hypercolumn and the motor time distribution) generalize across different experimental conditions.

Biological plausibility of parameters

The agreement between the optimal model predictions and the data collected from our subjects suggests that the human visual system may be optimal in visual search. Our model uses $n_H = 16$ uncorrelated, orientation-tuning neurons per visual location, each with a half tuning width of 22° and a maximum firing rate (estimated from the subjects) of approximately 17Hz. The tuning width agrees with V1 physiology in primates [39]. While our model appears to have underestimated the maximum firing rate of cortical neurons, which ranges from 30Hz to 70Hz [39], and the population size n_H (which may be in the order of hundreds), actual V1 neurons are

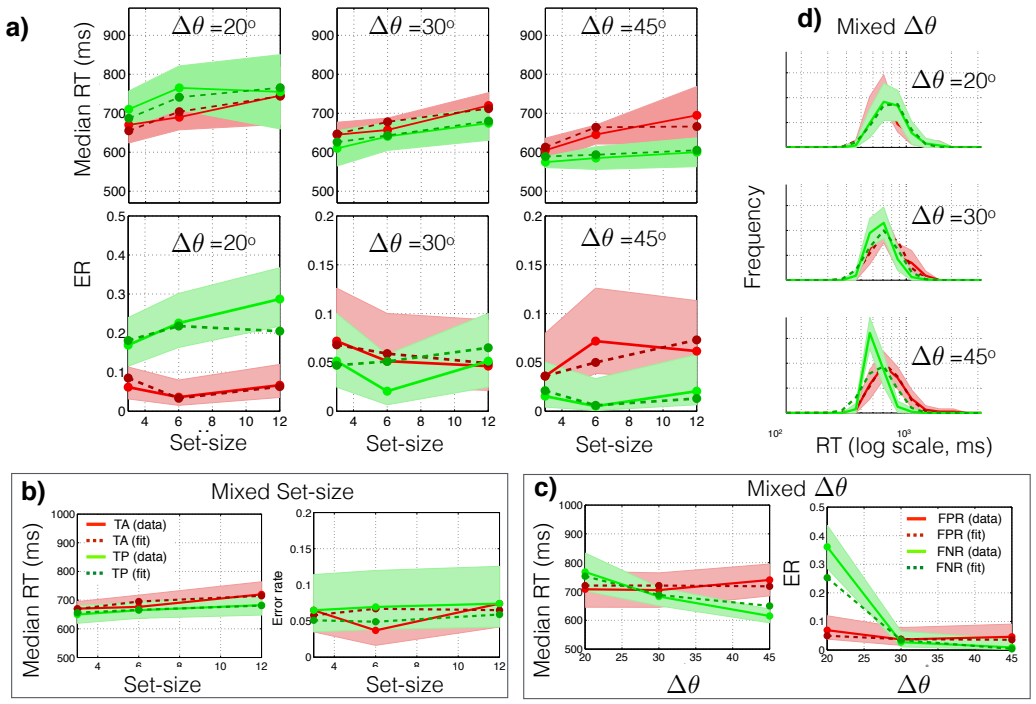

Figure 3.9: **Behavioral data of a randomly selected human subject and fits (ER, median RT and RT distributions) using SPRT**. (**a**) **Exp. 1**: "Blocked" design. All set-size M and orientation contrast $\Delta\theta$ combinations share the same hypercolumn and non-perceptual parameters; the decision thresholds are specific to each $\Delta\theta$-M pair. Fits are shown for RTs (first row) and ER (second row). (**b-c**) RT and ER for **Exp. 2**, the"mixed set size" (**b**) and **Exp. 3**, the "mixed contrast" design (**c**). (**d**) RT histogram for the "mixed contrast" design, grouped by orientation contrast.

correlated, hence the equivalent number of independent neurons is smaller than the measured number. For example, take a population of $n_H = 16$ independent Poisson neurons, all with a maximum firing rate of 17Hz, and combine every group of three of them into a new neuron. This will generate a population of 560 correlated neurons with a maximum firing rate of 51Hz and a correlation coefficient of 0.19, which is close to the experimentally measured average of 0.17 [39] (see [45] for a detailed discussion on the effect of sparseness and correlation between neurons). Therefore, our estimates of the model parameters are consistent with primate cortical parameters. The parameters of different subjects are close but not identical, matching the known variability within the human population [44], [46]. Finally, the fact that estimating model parameters from data collected in the blocked experiments allows the model to predict data collected in the mixed experiments does suggest that the model parameters mirror physiological parameters in our subjects.

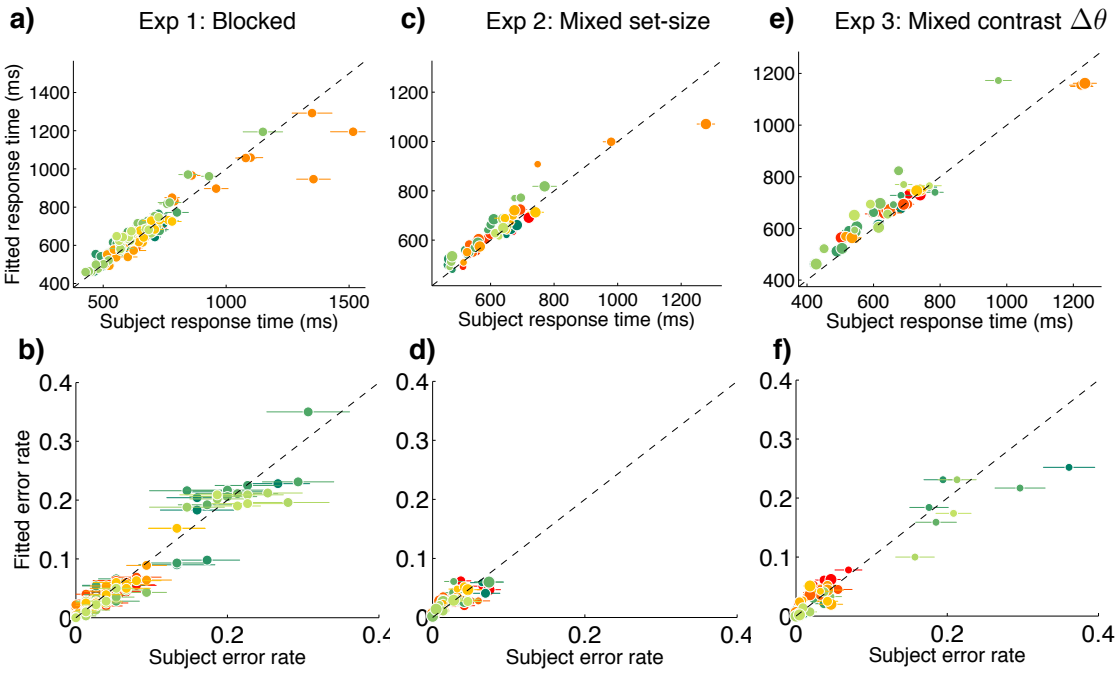

Figure 3.10: **Synopsis of fits to nine individual subjects.** The rows correspond respectively to three designs: **Exp. 1** (blocked), **Exp. 2** (mixed contrast) and **Exp. 3** (mixed set size). The maximum firing rate of the hypercolumn λ_{max} and the two non-decision parameters for each subject are fitted using only the blocked design experiment, and used to predict median RT and ER for the two mixed design experiments. Colors are specific to subject. The small, medium and large dots correspond, respectively, to the orientation contrast of 20°, 30°, and 45° in (**c-d**), and to the set-sizes 3, 6, and 12 in (**e-f**).

3.6 Spiking network implementation

Finally, we explore the physical realization of SPRT and show that a simple network of spiking neurons may implement a close approximation to the decision strategy.

Local log likelihoods

We first explain how to compute $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$, the local log likelihood of the stimulus taking on orientation θ, from spiking inputs $\boldsymbol{X}_{1:t}$ from V1. $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ is the building block of $S(\boldsymbol{X}_{1:t})$ (**Eq.** A.3). Consider one spatial location, the log likelihood is (derived in Appendix **Eq.** A.2):

$$\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) = \sum_{s=1}^{K_t} W_\theta^{i(s)} + \text{const.} \tag{3.13}$$

The first term is a diffusion, where each spike causes a jump in $\mathcal{L}_\theta$. Due to this

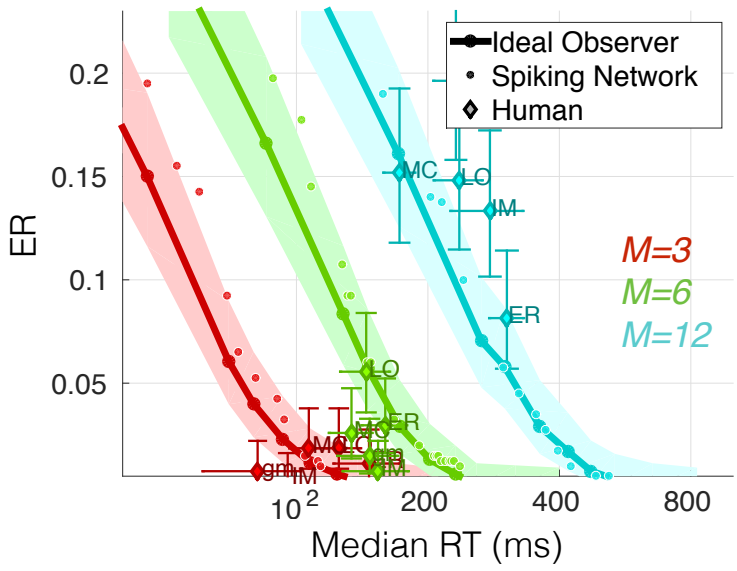

Figure 3.11: **Speed accuracy tradeoff.** ER vs RT tradeoff in the mixed set-size task (**Exp. 3**, **Fig. 3.2d**) of five human subjects (ER, IM gm, MC, and LO) with similar estimated internal parameters, as well as of SPRT (**Eq. 3.2**) and the spiking network implementation (**Sec. 3.6**) using the same internal parameters. The set-size takes value from $\{3, 6, 12\}$, and the orientation contrast is fixed at 30°.

property any linear combination of the diffusions, such as that of **Eq. 3.3**, is also a diffusion. This term can be implemented by integrate-and-fire [47] neurons, one for each relevant orientation $\theta \in \Theta_T \bigcup \Theta_D$, that receive afferent connections from all hypercolumn neurons with connection weights $w_\theta^i = \log \lambda_\theta^i$. The constant term is computationally irrelevant because it does not depend on the stimulus orientation θ; it may be removed by a gain-control mechanism to prevent the dynamic range of membrane potential from exceeding its physiological limits [48]. Specifically, one may subtract from each $\mathcal{L}_\theta$ a common quantity, e.g. the average value of the all the $\mathcal{L}_\theta$'s without changing $S(X_{1:t}^l)$ in **Eq. 3.11**.

Average gain-control Average gain-control is the process of subtracting the mean from the $\mathcal{L}_\theta$'s to remove unnecessary constants for decision and maintaining membrane potentials within physiological limits. Average gain-control may be conveniently done at the input using feedforward connections only. Specially, let $y_\theta(t)$ denote the mean-subtracted $\mathcal{L}_\theta$ signal, $w_\theta^i = \log \lambda_\theta^i$ denote the weights in **Eq. 3.13**, and $X_{i,t} \in \{0, 1\}$ denote the instantaneous firing event during time $(t-1)\Delta$ to $t\Delta$ from neuron i. The desired gain-controlled signal $y_\theta(t)$ may be computed by linear

integration, as shown in **Fig. 3.12a**:

$$\dot{y}_\theta(t) = \sum_i \left(w_\theta^i - \frac{\sum_{\theta'} w_{\theta'}^i}{n_H} \right) X_{i,t}. \tag{3.14}$$

Signal Transduction

The log likelihood $\mathcal{L}_\theta$ must be transmitted downstream for further processing. However, $\mathcal{L}_\theta$ is a continuous quantity whereas the majority of neurons in the central nervous system are believed to communicate via action potentials. We explored whether this communication may be implemented using action potentials [49] emitted from an integrate-and-fire neuron. Consider a sender neuron communicating its membrane potential to a receiver neuron. The sender may emit an action potential whenever its membrane potential surpasses a threshold τ_s. After firing, the membrane potential drops to its resting value, and the sender enters a brief refractory period whose duration (about $1ms$) is assumed to be negligible (in our simulations, time is discretized into $\Delta = 1ms$ bins, so we can model the refractory period by enforcing the condition that at most one spike can happen per bin for the sender neuron). If the synaptic strength between the two neurons is also τ_s, the receiver may decode the signal by simply integrating such weighted action potentials over time. This coding scheme loses some information due to discretization. Varying the discretization threshold τ_s trades off the quality of transmission with the number of action potentials: a lower threshold will limit the information loss at the cost of producing more action potentials. Surprisingly, we find that the performance of the spiking network is very close to that of the Bayesian observer, even when τ_s is set high, so that a small number of action potentials is produced (see **Fig. 3.12d,f** for the quality of approximation for a toy signal and **Fig. 3.6a-c** for the quality of approximation for actual signals in SPRT). The network behavior is quite insensitive to τ_s, thus we do not consider τ_s as a free parameter, and set its value to $\tau_s = 0.5$ in our experiments.

Softmax

One of the fundamental computations in **Eq. 3.10** is the softmax function (**Eq. 3.6**). It requires taking exponentials and logarithms, which have not yet been shown to be within a neuron's repertoire. Fortunately, it has been proposed that softmax may be approximated by a simple maximum [29], [42], and implemented using a winner-take-all mechanism [50], [51] with spiking neurons [52]. Through numerical

experiments we find that this approximation results in almost no change to the network's behavior (see **Fig. 3.12e**). This suggests that an exact implementation of softmax is not critical, and other mechanisms that may be more neurally plausible have similar performances.

One common implementation [50] of the softmax is described as follows. For a set of n_H spiking neurons, let $X_{i,t} \in \{0, 1\}$ denote whether neuron i has spiked in time $((t-1)\Delta, t\Delta]$. We introduce an additional n_H neurons $\{y_i\}_{i=1}^{n_H}$, where $y_i(t)$ denotes the membrane potential of the i-th additional neuron at time t. The desired quantity is the softmax over the cumulative signal in $\boldsymbol{X}_{1:t}$, denoted by $z(t)$. In other words

$$z(t) \triangleq \underset{i=1,\ldots,n_H}{\mathcal{S}\max} \left(w_i \sum_{t'=1}^{t} X_{i,t'} \right),$$

and $z(t)$ may be approximated by $\tilde{z}(t)$ using the following neuron equations (derived from Taylor expansion):

$$\dot{\tilde{z}}(t) = \sum_i y_i(t) w_i X_{i,t}, \tag{3.15}$$

$$\dot{y_i}(t) = y_i(t)(w_i X_{i,t} - \dot{\tilde{z}}(t)). \tag{3.16}$$

Fig. 3.12e shows that $\tilde{z}(t)$ approximates $z(t)$ well in a simple setup of seven neurons with a common and small incoming weights $w_i = 0.05$ across all neurons.

The time it takes for the winner-take-all network to converge is typically small (on the ms level for tens of neurons, scaling logarithmically with the number of neurons [50]) compared to the inter-spike-intervals of the input neurons (around $30ms$ per neuron, and $12ms$ for a hypercolumn of $n_H = 16$ neurons per visual location [45]).

Decision

Finally, the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ is compared to a pair of thresholds to reach a decision (**Eq. 3.2**). The positive and negative parts of $S(\boldsymbol{X}_{1:t})$, $(S(\boldsymbol{X}_{1:t}))^+$ and $(-S(\boldsymbol{X}_{1:t}))^+$, may be represented separately by two mutually inhibiting neurons [53], where $(\cdot)^+$ denotes halfwave-rectification: $(x)^+ \triangleq \max(0, x)$. We can implement **Eq. 3.2** by simply setting the firing thresholds of these neurons to the decision thresholds τ_1 and $-\tau_0$ respectively.

Alternatively, $S(\boldsymbol{X}_{1:t})$ may be computed by a mechanism akin to the ramping neural activity observed in decision-implicated areas such as the frontal eye field [19]–[21].

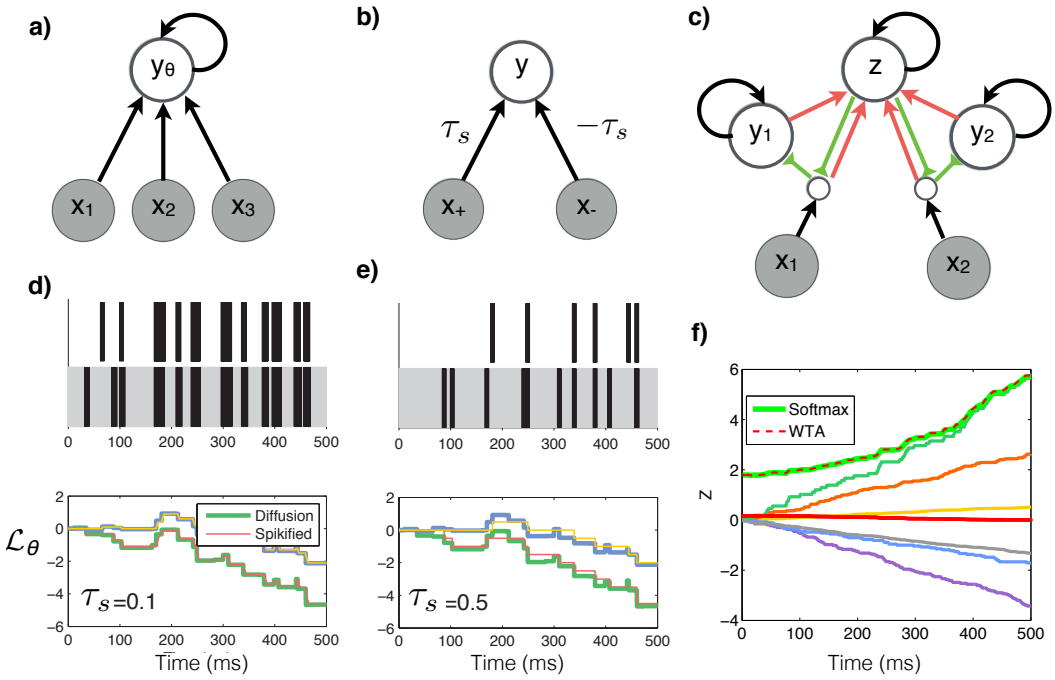

Figure 3.12: **Spiking implementation** (**a**) A feedforward network implemented the average gain-controlled network of **Eq. 3.14**. (**b**) Signal transduction. The positive and negative parts of the signal in x are encoded with integrate-and-fire neurons and transmitted to the receiver neuron y. (**c**) Winner-take-all circuit for computing the softmax (**Eq. 3.15** and **Eq. 3.16**). (**d-e**, Top) Two sender neurons communicate their membrane potentials using spike trains to a receiver neuron (only the negative neurons are shown). (**d-e**, Bottom) The receiver reassembles the spike trains (thick lines) and reconstructs the senders' membrane potentials (thin lines). (**f**) Comparison between the ground-truth and the WTA implementation of softmax of seven neurons over time.

$(S(\boldsymbol{X}_{1:t}))^+$ and $(-S(\boldsymbol{X}_{1:t}))^+$ could be converted to two trains of action potentials using the same encoding scheme described above in the Signal Transduction section. The resultant spike trains may be the input signal of an accumulator model (e.g. [16]). The model has been shown to be implementable as a biophysically realistic recurrent network [23], [54], [55] and capable of producing and thresholding ramping neural activity to trigger motor responses [19]–[22], [56]. While both neural implementations of $S(\boldsymbol{X}_{1:t})$ are viable options, in the simulations used in this study we opted for the first.

Network structure

If we combine the mechanisms discussed above, i.e. local gain-control, an approximation of softmax, a spike-based coding of analog log likelihood values as well as the decision mechanism, we see that the mathematical computations required by the SPRT can be implemented by a deep recurrent network of spiking neurons (**Fig. 3.5a**).

The overall network structure is identical to the diagram (**Fig. 3.5b**). It is composed of local "hypercolumn readout" networks (**Fig. 3.5b**), and a central circuit that aggregates information over the visual field. The local network computes the local log likelihood ratio $S^l(\boldsymbol{X}^l_{1:t})$ (**Eq. 3.11**) and simultaneously computes the local log likelihood for each CDD. The CDD log likelihoods are aggregated over all locations and sent to a gain-control unit to estimate the posterior of the CDD, $Q_\phi = \log P(\phi|\boldsymbol{X}_{1:t})$, which captures the most likely set-size and orientation contrast. At each time instant this estimate is fed back to the local networks to compute $S(\boldsymbol{X}^l_{1:t})$ (**Eq. 3.11**).

It is important to note that both the **structure** and **the synaptic weights** of the visual search network described above were derived **analytically** from the hypercolumn parameters (the shape of the orientation-tuning curves), the decision thresholds, and the probabilistic description of the task. The network designed for heterogeneous visual search could dynamically switch to simpler tasks by adjusting its priors (e.g. $P(\phi)$). The network has only three degrees of freedom, rather than a large number of network parameters [29], [57].

As shown in **Fig. 3.11,** the spiking implementation approximates SPRT very well, indicating that the brain *can* implement optimal Bayesian sequential reasoning using simple neural mechanisms.

3.7 Chapter summary

Searching for objects amongst clutter is one of the most valuable functions of our sensory systems. Best performance is achieved with fast response time (RT) and low error rates (ER); however, response time and error rates are competing requirements which have to be traded off against each other. The faster one wishes to respond, the more errors one makes due to the limited rate at which information flows through the senses. Conversely, if one wishes to reduce error rates, decision times become longer. In order to study the nature of this trade-off we derived SPRT for visual search; the input signal to the model is action potentials from orientation-selective

hypercolumn neurons in primate striate cortex V1, the output of the model is a binary decision (target-present versus target-absent) and a decision time.

Five free parameters uniquely characterize the model: the maximum firing rate of the input neurons and the maximum tolerable false-alarm and false-reject error rates, as well as two parameters characterizing response delays that are unrelated to decision. Once these parameters are set, RT histograms and ER may be computed for any experimental condition. Our model may be implemented by a deep neural network composed of integrate-and-fire and winner-take-all mechanisms. The network structure is completely deterministic given the probabilistic structure of the search task. Signals propagate from layer to layer mostly in a feed-forward fashion; however, we find that two feedback mechanisms are necessary: (i) gain control (lateral inhibition) that is local to each hypercolumn and has the function of maintaining signals within a small dynamic range, and (ii) global inhibition that estimates the complexity of the scene. Qualitative comparison of model predictions with human behavior suggests that the visual system of human observers indeed does estimate scene complexity as it carries out visual search, and that this estimate is used to control the gain of decision mechanisms.

Despite the parsimony, our model is able to quantitatively predict human behavior in a variety of visual search conditions. Without physiological measurements of the hypercolumn parameters (number of neurons, maximum firing rate, etc) directly from human subjects, one can not assess optimality. After all, we may be overestimating the signal-to-noise ratio in the front-end while humans are sub-optimal. Nonetheless, the estimated hypercolumn parameters are plausible, suggesting that humans may employ an optimal strategy for visual search.

References

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[2] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, no. 3, pp. 433–458, Jul. 1989, ISSN: 0033-295X.

[3] P. Verghese and K. Nakayama, "Stimulus discriminability in visual search," *Vision Research*, vol. 34, no. 18, pp. 2453–2467, 1994.

[4] J. Palmer, "Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks," *Vision Research*, vol. 34, no. 13, pp. 1703–1721, 1994.

[5] M. Carrasco and Y. Yeshurun, "The contribution of covert attention to the set-size and eccentricity effects in visual search.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 2, pp. 673–692, 1998.

[6] E. L. Cameron, J. C. Tai, M. P. Eckstein, and M. Carrasco, "Signal detection theory applied to three visual search tasks-identification, yes/no detection and localization," *Spatial Vision*, vol. 17, no. 4, pp. 295–326, 2004.

[7] J. M. Wolfe, T. S. Horowitz, and N. M. Kenner, "Rare items often missed in visual searches," *Nature*, vol. 435, no. 7041, pp. 439–440, 2005.

[8] V. Navalpakkam, C. Koch, A. Rangel, and P. Perona, "Optimal reward harvesting in complex perceptual environments," *Proceedings of the National Academy of Sciences*, vol. 107, no. 11, pp. 5232–5237, 2010.

[9] J. M. Wolfe, E. M. Palmer, and T. S. Horowitz, "Reaction time distributions constrain models of visual search," *Vision Research*, vol. 50, no. 14, pp. 1304–1311, 2010.

[10] M. P. Eckstein, "Visual search: A retrospective," *Journal of Vision*, vol. 11, no. 5, p. 14, 2011.

[11] M. Pomplun, T. W. Garaas, and M. Carrasco, "The effects of task difficulty on visual search strategy in virtual 3d displays," *Journal of Vision*, vol. 13, no. 3, p. 24, 2013.

[12] R. Ratcliff, "Theoretical interpretations of the speed and accuracy of positive and negative responses.," *Psychological Review*, vol. 92, no. 2, p. 212, 1985.

[13] J. R. Busemeyer and J. T. Townsend, "Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment.," *Psychological Review*, vol. 100, no. 3, pp. 432–459, 1993.

[14] M. Usher and J. L. McClelland, "The time course of perceptual choice: The leaky, competing accumulator model.," *Psychological Review*, vol. 108, no. 3, p. 550, 2001.

[15] M. Shadlen and W. Newsome, "Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey," *Journal of Neurophysiology*, vol. 86, no. 4, pp. 1916–1936, 2001.

[16] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks.," *Psychological Review*, vol. 113, no. 4, p. 700, 2006.

[17] S. D. Brown and A. Heathcote, "The simplest complete model of choice response time: Linear ballistic accumulation," *Cognitive Psychology*, vol. 57, no. 3, pp. 153–178, 2008.

[18] J. M. Wolfe, "Guided search 4.0," *Integrated Models of Cognitive Systems*, pp. 99–119, 2007.

[19] B. A. Purcell, J. D. Schall, G. D. Logan, and T. J. Palmeri, "From salience to saccades: Multiple-alternative gated stochastic accumulator model of visual search," *The Journal of Neuroscience*, vol. 32, no. 10, pp. 3433–3446, 2012.

[20] G. F. Woodman, M.-S. Kang, K. Thompson, and J. D. Schall, "The effect of visual search efficiency on response preparation neurophysiological evidence for discrete flow," *Psychological Science*, vol. 19, no. 2, pp. 128–136, 2008.

[21] R. P. Heitz and J. D. Schall, "Neural mechanisms of speed-accuracy tradeoff," *Neuron*, vol. 76, no. 3, pp. 616–628, 2012.

[22] M. E. Mazurek, J. D. Roitman, J. Ditterich, and M. N. Shadlen, "A role for neural integrators in perceptual decision making," *Cerebral Cortex*, vol. 13, no. 11, pp. 1257–1269, 2003.

[23] K.-F. Wong, A. C. Huk, M. N. Shadlen, and X.-J. Wang, "Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making," *Frontiers in Computational Neuroscience*, vol. 1, 2007.

[24] J. Palmer, A. C. Huk, and M. N. Shadlen, "The effect of stimulus strength on the speed and accuracy of a perceptual decision," *Journal of Vision*, vol. 5, no. 5, pp. 376–404, 2005.

[25] J. Drugowitsch, R. Moreno-Bote, A. K. Churchland, M. N. Shadlen, and A. Pouget, "The cost of accumulating evidence in perceptual decision making," *The Journal of Neuroscience*, vol. 32, no. 11, pp. 3612–3628, 2012.

[26] W. S. Geisler, "Sequential ideal-observer analysis of visual discriminations.," *Psychological Review*, vol. 96, no. 2, pp. 267–314, 1989.

[27] J. Palmer, P. Verghese, and M. Pavel, "The psychophysics of visual search," *Vision Research*, vol. 40, no. 10, pp. 1227–1268, 2000.

[28] P. Verghese, "Visual search and attention: A signal detection theory approach," *Neuron*, vol. 31, no. 4, pp. 523–535, 2001.

[29] W. J. Ma, V. Navalpakkam, J. M. Beck, R. Van Den Berg, and A. Pouget, "Behavior and neural basis of near-optimal visual search," *Nature Neuroscience*, vol. 14, no. 6, pp. 783–790, 2011.

[30] W. S. Geisler, "Contributions of ideal observer theory to vision research," *Vision Research*, vol. 51, no. 7, pp. 771–781, 2011.

[31] S. S. Shimozaki, W. A. Schoonveld, and M. P. Eckstein, "A unified bayesian observer analysis for set size and cueing effects on perceptual decisions and saccades," *Journal of Vision*, vol. 12, no. 6, p. 27, 2012.

[32] D. Green and J. Swets, *Signal detection theory and psychophysics*. Peninsula, Los Altos, CA, 1966.

[33] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[34] D. J. Felleman and D. C. Van Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[35] T. D. Sanger, "Probability density estimation for the interpretation of neural population codes," *Journal of Neurophysiology*, vol. 76, no. 4, pp. 2790–2793, 1996.

[36] E. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Network: Computation in Neural Systems*, vol. 12, no. 2, pp. 199–213, 2001.

[37] E. P. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz, "Characterization of neural responses with stochastic stimuli," *The Cognitive Neurosciences*, vol. 3, pp. 327–338, 2004.

[38] J. Beck, W. Ma, R. Kiani, T. Hanks, A. Churchland, J. Roitman, M. Shadlen, P. Latham, and A. Pouget, "Probabilistic population codes for bayesian decision making," *Neuron*, vol. 60, no. 6, pp. 1142–1152, 2008, ISSN: 0896-6273.

[39] A. B. Graf, A. Kohn, M. Jazayeri, and J. A. Movshon, "Decoding the activity of neuronal populations in macaque primary visual cortex," *Nature Neuroscience*, vol. 14, no. 2, pp. 239–245, 2011.

[40] R. L. Goris, J. A. Movshon, and E. P. Simoncelli, "Partitioning neuronal variability," *Nature Neuroscience*, vol. 17, no. 6, pp. 858–865, 2014.

[41] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[42] B. Chen, V. Navalpakkam, and P. Perona, "Predicting response time and error rates in visual search," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

[43] M. Stone, "Models for choice-reaction time," *Psychometrika*, vol. 25, no. 3, pp. 251–260, 1960.

[44] E. Palmer, T. Horowitz, A. Torralba, and J. Wolfe, "What are the shapes of response time distributions in visual search?" *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 1, pp. 58–71, 2011.

[45] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[46] D. C. Van Essen, W. T. Newsome, and J. H. Maunsell, "The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability," *Vision Research*, vol. 24, no. 5, pp. 429–448, 1984.

[47] P. Dayan and L. Abbott, "Theoretical neuroscience: Computational and mathematical modeling of neural systems," *Journal of Cognitive Neuroscience*, vol. 15, no. 1, pp. 154–155, 2003.

[48] M. Carandini, D. J. Heeger, and J. A. Movshon, "Linearity and gain control in v1 simple cells," in *Models of Cortical Circuits*, Springer, 1999, pp. 401–443.

[49] C. M. Gray and D. A. McCormick, "Chattering cells: Superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex," *Science*, vol. 274, no. 5284, pp. 109–113, 1996.

[50] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*, Springer, 1987, pp. 115–141.

[51] H. S. Seung, "Reading the book of memory: Sparse sampling versus dense mapping of connectomes," *Neuron*, vol. 62, no. 1, pp. 17–29, 2009.

[52] M. Oster, R. Douglas, and S.-C. Liu, "Computation with spikes in a winner-take-all network," *Neural Computation*, vol. 21, no. 9, pp. 2437–2465, 2009.

[53] F. Gabbiani and C. Koch, "Coding of time-varying signals in spike trains of integrate-and-fire neurons with random threshold," *Neural Computation*, vol. 8, no. 1, pp. 44–66, 1996.

[54] X.-J. Wang, "Probabilistic decision making by slow reverberation in cortical circuits," *Neuron*, vol. 36, no. 5, pp. 955–968, 2002.

[55] C.-C. Lo and X.-J. Wang, "Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks," *Nature Neuroscience*, vol. 9, no. 7, pp. 956–963, 2006.

[56] P. Cassey, A. Heathcote, and S. D. Brown, "Brain and behavior in decision-making," *PLoS Computational Biology*, vol. 10, no. 7, e1003700, 2014.

[57] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.

[1] B. Chen and P. Perona, "Speed versus accuracy in visual search: Optimal performance and neural architecture," *Journal of Vision*, vol. 15, no. 16, pp. 9–9, 2015.

Chapter 4

SCOTOPIC VISUAL RECOGNITION

Sequential Reasoning without the Probabilistic Model

Our second project is scotopic visual recognition, which aims to recognize objects with as little light as possible. This project is motivated by real-world applications ranging from biological imaging to astrophysics. Unlike visual search (**Ch. 3**), most practical vision applications do not have the luxury of knowing the full probabilistic model for the task at hand. To circumvent this problem we proposed techniques to train a sequential algorithm directly to optimize the speed versus accuracy tradeoff (SAT).

4.1 Motivations

Just like biological systems, computer vision systems are optimized for accuracy and speed. Accuracy is well understood as the success rate at identifying object classes, estimating object poses, etc. Speed depends on the time it takes to capture an image (exposure time) and the time it takes to compute the answer. Computer vision researchers typically assume that there is plenty of light and a large number of photons may be collected very quickly, thus speed is limited by computation. This is called *photopic vision* where the image, while difficult to interpret, is (almost) noiseless; researchers ignore exposure time and focus on the trade-off between accuracy and computation time (e.g. Fig 10 of [1]).

> In images with eight bits per pixel of signal (i.e. SNR=256), pixels collect 10^4 – 10^5 photons [2]. In full sunlight the exposure time is about 1/1000 s which is negligible compared to typical computation times.

Consider now the opposite situation, which we call *scotopic vision*, where photons are few and precious, and exposure time is long compared to computation time. As computation time becomes a small additive constant, the design tradeoff is between accuracy and exposure time [3]. There are multiple situations where trading off accuracy with exposure time is compelling. (1) One may be trying to sense/control dynamics that are faster than the exposure time that guarantees good quality pictures, e.g. automobiles and quadcopters [4]. (2) In competitive scenarios, such as sports, a fraction of a second may make all the difference between defeat

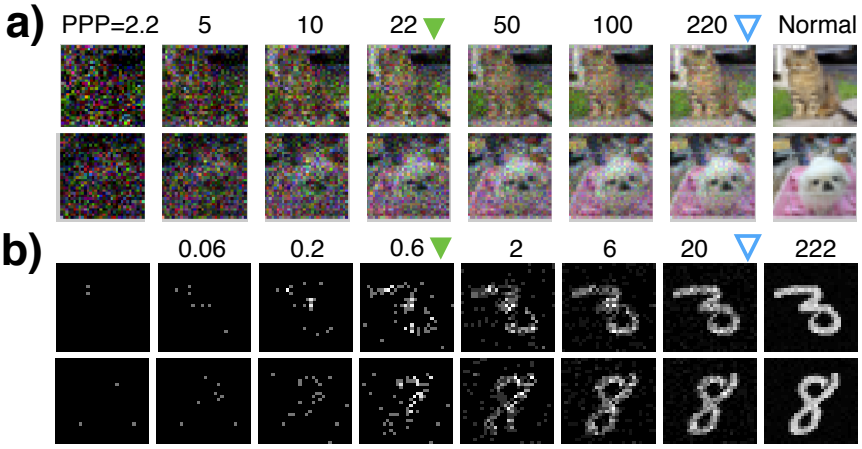

Figure 4.1: **Classification with few photons and speed-accuracy tradeoff.** Cumulative photon count N_t generated using sample images from the **(a)** CIFAR10 dataset and the **(b)** MNIST dataset with increasing average photons per pixel (PPP). PPP is proportional to the exposure time t. The images were obtained by simulating photon arrival times (**Sec. A.2**). Blue hollow arrows indicate the median PPP required for our scotopic classifier (WaldNet) to achieve comparable error rates (21%) as the model trained and tested using images under normal lighting conditions with about $2^7 \approx 10^4$ PPP (see **Sec. A.2** for protocol). Considerable speedups, of about two orders of magnitude, may be obtained by making classification happen as soon as a sufficient number of photons has been collected. Considerable further speed gains may be achieved by trading-off classification performance with decision speed: green solid arrows indicate the median PPP required to to maintain error rates below 22% for CIFAR and 1% for MNIST.

and victory [5]. (3) Sometimes prolonged imaging has negative consequences, e.g. because phototoxicity and bleaching alter a biological sample [6] or because of health risks in medical imaging [7]. (4) In sensor design, reduced photon counts allow for imaging with smaller pixels and ultra-high resolution [8], [9]. (5) Sometimes there is little light in the environment, e.g. at night, and obtaining a good quality image takes a long time relative to achievable computational speed. Thus, it is compelling to understand how many photons are needed for good-enough vision, and how one can make visual decisions as soon as a sufficient number of photons has been collected.

> The term 'scotopic / photopic vision' literally means 'vision in the dark / with plenty of light'. It is usually associated to the physiological state where only rods, not cones, are active in the retina. We use 'scotopic vision' to denote the general situation where a visual system is starved for photons, regardless of the technology used to capture the image.

Our work is further motivated by the recent development of *photon-counting imaging sensors*: single photon avalanche diode arrays [10], quanta image sensors [9], and gigavision cameras [8]. Instead of returning a high-quality image after a fixed exposure time, these sensors detect and report single photon arrival events at high frequencies. This ability to manipulate photon acquisition with fine granularity makes photon-counting sensors ideal for scotopic vision applications. Current computer vision technology has not yet taken advantage of these sensors.

4.2 Contributions

While scotopic vision has been studied in the context of the physiology and technology of image sensing [11], [12], as well as the physiology and psychophysics of visual discrimination [13] and visual search [14], little is known regarding the computational principles for high-level visual tasks, such as categorization and detection, in scotopic settings. Prior work on photon-limited image classification [15] deals with a single image, and does not study the trade-off between exposure time and accuracy. Instead, our work explore scotopic visual categorization on modern datasets such as MNIST and CIFAR10 [16], [17].

Sequential testing has appeared in the computer vision literature [18]–[20] in order to *shorten computation time*. These algorithms assume that all visual information ('the image') is present at the beginning of computation, thus focus on reducing computation time in photopic vision. By contrast, *our work aims to reduce capture time* and is based on the assumption that computation time is negligible when compared to image capture time. The similarity between the two lines of work is therefore only superficial.

Our main contributions are:
1. We present a **computational framework** for scotopic classification that dynamically decides the image exposure time for SAT.
2. When a probabilistic model of the classification task is given, we design a feed-forward architecture yielding **any-time, quasi-optimal** scotopic classification.
3. When the probabilistic model is not available, we propose a **learning algorithm** to train the architecture for optimizing the SAT.
4. We conduct a **robustness analysis** with respect to sensor noise in current photon-counting sensor prototypes.

4.3 Framework for scotopic classification

Quantized sensory input

Our computational framework starts from a model of the sensory input. Each pixel in an image reports the brightness estimate of a cone of visual space by counting photons coming from that direction. The estimate improves over time.

To begin we consider a simpler version of the problem where the assumptions (**Ch. 2**) are met for SPRT. We assume that 1) the world is stationary during the imaging process (this may be justified as many photon-counting sensors sample the world at $> 1kHz$ [8], [9]); 2) photon arrival times follow a homogeneous Poisson process (details below) and 3) a probabilistic classifier based on photon counts is available. Assumption 3) may not be satisfied for practical object recognition classification problems, therefore we discuss how to do without this assumption in **Sec. 4.3**.

Poisson noise model

Sensors are corrupted by several intrinsic noise sources [21]. **Shot noise**: the number of photons incident on a pixel i in the t-th time interval, $X_{t,i}$, follows a Poisson distribution whose rate λ_i (Hz) depends on both the pixel intensity $I_i \in [0, 1]$ and a **dark current** ϵ_{dc}:

$$P(X_{t,i} = k) = Poisson(k|\lambda_i t\Delta) = Poisson(k|\lambda^\phi \frac{I_i + \epsilon_{dc}}{1 + \epsilon_{dc}} t\Delta), \tag{4.1}$$

where λ^ϕ is the illuminance (maximum photon count per pixel) per unit time [2], [8], [21], [22]. During readout, the photon count is additionally corrupted first by the amplifier's **read noise**, which is an additive Gaussian, then by the **fixed-pattern noise** which may be thought of as a multiplicative Gaussian noise [23]. As photon-counting sensors are designed to have low read noise and low fixed pattern noise[9], [10], [22], we focus on modeling the shot noise and dark current only. We will show (**Sec. 4.4**) that our models are robust against all four noise sources.

> According to the stationary assumption there is no need to model *motion-induced blur*. Additionally, for simplicity we do not model *charge bleeding and cross-talk* in colored images, and assume that they will be mitigated by the sensor community [24].

When the illuminance λ^ϕ of the environment is fixed, the amount of photons is roughly linear in the exposure time t (**Eq. 4.1**). Hence we use the number of photons

per bright pixel (PPP) interchangeably with the exposure time t. i.e.:

$$PPP = \lambda^{\phi} t \Delta. \tag{4.2}$$

PPP= 1 means that a pixel with maximum intensity has collected 1 photon. Since the information content in the image is directly related to the number of photons, from now on we measure response time in terms of PPP instead of exposure time. **Fig. 4.1** shows a series of images from the CIFAR10 dataset [16] with increasing PPP.

Sequential probability ratio test for scotopic classification

Assume that a probabilistic model is available to interpret the sensory input given the class label – either provided by the application or learned from labeled data using techniques described in **Sec. 4.3** – we can apply SPRT to classify the photon streams. Since the classification task may contain multiple categories, the SPRT formulation **Eq. 2.3** needs to be extended to handle multiple hypothesis testing [25], [26].

Let $S_c(\boldsymbol{X}_{1:t}) \triangleq \log \frac{P(C=c|\boldsymbol{X}_{1:t})}{P(C\neq c|\boldsymbol{X}_{1:t})}$ denote the class posterior probability ratio of the visual category C for photon count input $\boldsymbol{X}_{1:t}, \forall c \in \{1, \ldots, K\}$, and let τ be an appropriately chosen threshold. SPRT conducts a simple accumulation-to-threshold procedure to estimate the category $\hat{C}$:

$$\begin{aligned} &\text{Compute } c^* = \arg\max_{c=1,\ldots,K} S_c(\boldsymbol{X}_{1:t}) \\ &\text{if } S_{c^*}(\boldsymbol{X}_{1:t}) > \tau : \text{ report } \hat{C} = c^* \\ &\text{otherwise } : \text{increase exposure time } t. \end{aligned} \tag{4.3}$$

Static versus dynamic exposure time models

In essence, SPRT decides when to respond dynamically, based on the stream of observations accumulated so far. As a result of the trial-by-trial variation of the signal, the response time also varies trial by trial. This regime is called " **free-response**" (FR), in contrast to the " **interrogation**" (INT) regime, typical of photopic vision, where a fixed-length observation is collected for each trial [27]. The observation length may be chosen according to a training set and fixed a priori. In both regimes, the length of observation should take into account the cost of errors, the cost of time, and the difficulty of the classification task.

Despite the striking similarity between the two regimes, SPRT (the FR regime) outperforms the INT regime, as we prove here for the case where the observations are i.i.d., and demonstrate empirically in **Sec. 4.4**.

Theorem 1 ***Free-response is asymptotically better than interrogation***. *Assume that a probabilistic model is given to compute $S(\boldsymbol{X}_{1:t})$, and $\boldsymbol{X}_t$ is i.i.d. in time. Consider an FR algorithm that runs SPRT on $S(\boldsymbol{X}_{1:t})$ and let ϵ_{FR} and T_{FR} be its error rate and stochastic decision time. Also consider an INT algorithm with a fixed-length observation of t_{INT} that achieves an error of ϵ_{INT}. We have that the Bayes risk (**Eq. 2.1**) of the FR algorithm is less than or equal to that of the INT algorithm. In other words, as $\eta \to 0$:*

$$\mathbb{E}[T_{FR}] + \eta\epsilon_{FR} \leq t_{INT} + \eta\epsilon_{INT}.$$

Proof We prove the statement for binary classification with equal prior ($K = 2$, **Eq. 2.3**, the proof extends trivially to larger K). Consider all $\boldsymbol{X}_{1:t}$ generated from the positive class $C = 1$. Given an error rate requirement ϵ_{FR}, the FR algorithm sets up its threshold τ such that all the trials that terminate with $\hat{C} = 1$ must achieve a posterior probability of $1 - \epsilon_{FR}$, i.e. $P(C = 1|\boldsymbol{X}_{1:t}) = 1 - \epsilon_{FR}$, where $P(C = 1|\boldsymbol{X}_{1:t}) = Sigm(S(\boldsymbol{X}_{1:t}))$. Therefore, the threshold satisfies $Sigm(\tau) = 1 - \epsilon_{FR}$.

Since $\boldsymbol{X}_t$ is i.i.d. in time, $S(\boldsymbol{X}_{1:t}) = \sum_t S(X_t)$. Let $\mu \triangleq \mathbb{E}[X_t], \forall t$ represent the mean evidence accumulation rate (constant over time). The expected run time for the FR algorithm is

$$t_{FR} = \mathbb{E}[T_{FR}] = \frac{\tau}{\mu}.$$

Now consider an INT algorithm *with the same observation time* as the expected observation time for the FR algorithm, i.e. $t_{INT} = t_{FR}$. As $\eta \to 0$, $\epsilon_{FR} \to 0$, $t_{FR} \to \infty$ and $S(\boldsymbol{X}_{1:t_{FR}}) \geq 0$, a.s.. The error rate of the INT algorithm is

$$\begin{aligned} 1 - \epsilon_{INT} &= \mathbb{E}[Sigm(S(\boldsymbol{X}_{1:t_{FR}}))] \leq Sigm(\mathbb{E}[S(\boldsymbol{X}_{1:t_{FR}})]), a.s. \\ &= Sigm(\mu t_{FR}) = Sigm(\mu\frac{\tau}{\mu}) = Sigm(\tau) = 1 - \epsilon_{FR}, \end{aligned}$$

as a result of Jensen's inequality used on $Sigm(x)$, which is **concave** when $x \geq 0$.

Therefore as $\eta \to 0$, for any $t_{FR} = t_{INT}$, we have $\epsilon_{FR} \leq \epsilon_{INT}, a.s.$. Therefore for any pair of $\{t_{INT}, \epsilon_{INT}\}$ that minimizes Bayes risk for the INT algorithm, we can find an FR algorithm with $\{t_{FR}, \epsilon_{FR}\}$ that achieves a lower or equal Bayes risk. ∎

Computing class probabilities over time

The challenge of applying SPRT is to compute $S_c(\boldsymbol{X}_{1:t})$ for class c and the input stream $\boldsymbol{X}_{1:t}$ of variable exposure time t, or in a more information-relevant unit, variable PPP levels. Thanks to the Poisson noise model (**Eq. 4.1**), the sufficient statistics for observation $\boldsymbol{X}_{1:t}$ is the cumulative count $\boldsymbol{N}_t = \sum_{t'=1}^{t} \boldsymbol{X}_{t'}$ (visualized in **Fig. 4.1**), therefore we may rewrite $S_c(\boldsymbol{X}_{1:t})$ as $S_c(\boldsymbol{N}_t)$. It is evident that counts at different PPPs have different statistics. It would appear that a specialized system is required for each PPP level. This leads to the naive *ensemble* approach. Instead, we also propose a network called *WaldNet* that can process images at all PPPs and has the size of only a single specialized system. We describe the two approaches below.

> We insist on the need to distinguish between the cumulative count $\boldsymbol{N}_t$ and the conventional image, which is obtained by normalizing $\boldsymbol{N}_t$ to intensities within $[0, 255]$. By retaining the magnitude of the counts, $\boldsymbol{N}_t$ carries the uncertainty of the intensity estimates, which is crucial for evaluating the confidence of the class prediction.

A naïve approach: network ensembles

The simple idea is to build a separate model $S(\boldsymbol{N}_t)$ for the cumulative counts for each exposure time t (or light level PPP), either based on domain knowledge or learned from a training set. For best results one needs to select a list of representative light levels, and then apply each to input streams that were captured at the corresponding light level. For cumulative counts $\boldsymbol{N}_{t'}$ captured at light levels that are not on the list, one may simply apply the model with the closest light level. We refer to this as the 'ensemble' predictor.

One potential drawback of this ensemble approach is that training and storing multiple systems is *wasteful*. At different light levels, while the cumulative counts change drastically, the underlying statistical structure of the task stays the same. An approach that takes advantage of this relationship may lead to more parsimonious algorithms.

Model-based approach: WaldNet

An alternative is to exploit the knowledge about the cumulative counts across light levels. The variation in the input $\boldsymbol{N}_t$ has two independent sources: one is the stochasticity in the photon arrival times, and the other the intra- and inter- class variation of the real intensity values of the object. SPRT excels at reasoning about the first noise source while deep networks are ideal for capturing the second. Therefore

we propose *WaldNet*, a deep network for speed-accuracy tradeoff (**Fig. 4.2b-c**) that combines deep networks with SPRT. Standard deep networks such as convolutional networks [17] (ConvNets) can not be applied directly as their inputs all have an identical exposure time T (e.g. $T \approx 33ms$ in normal lighting conditions). Instead, WaldNet utilizes lowlight noise statistics (**Sec. 4.4**) to adjust the computation within a deep network over exposure time t in order to compute the log class probability ratios $S_c(\boldsymbol{N}_t)$ over time t.

We first assume that a *generative* model for the cumulative counts $\boldsymbol{N}_T$ is available, and use it to develop a generative model for WaldNet. Then we provide a *discriminative* model with the identical computational form as the generative model, which may be learned directly from data.

> The generative model is rather technical. Readers who are not familiar with the literature on restricted Boltzmann machines and deep belief networks [28], [29] are encouraged to skip directly to the next section that discusses the discriminative training of WaldNet.

We assume that the generative model of input photon counts takes the form of a deep belief network [29]. The deep belief network is composed of multiple stacks. A stack on layer l consists of an input vector $\boldsymbol{v}^{(l)}$, a hidden vector $\boldsymbol{h}^{(l)} \in \{0,1\}^{n_H^l}$ and a pooling vector $\boldsymbol{m}^{(l)} \in \{0,1\}^{n_M^l}$. The log posterior ratio of the pooling vector of one layer becomes the input vector of the layer above, $v_i^{l+1} = \log \frac{P(m_i^{(l)}=1)}{P(m_i^{(l)}=0)}$, and the last pooling vector encodes desired log class posterior ratio $S(\boldsymbol{N}_T)$. $\boldsymbol{m}^{(l)}$, $\boldsymbol{h}^l$ and $\boldsymbol{v}^{(l)}$ are connected convolutionally as in a ConvNet, as follows:

1. Each pooling unit $m_k^{(l)}$ oversees a non-overlapping group $G_k^{(l)}$ of hidden units where at most one hidden unit is allowed to be on. $m_k^{(l)} = 1$ represents the presence of an image feature (say a 45° edge) anywhere within a spatial neighborhood $G_k^{(l)}$ of the image, and $h_j^{(l)} = 1$ indicates that the feature's location is j. This formulation is a generalization of probabilistic max pooling [30].

2. Each hidden unit $h_k^{(l)}$ connects to a small (say 5×5) neighborhood of input units $\boldsymbol{v}^{(l)}$. For layers $l > 1$ the hidden-input relationship is a standard RBM [28], [30], [31]. In the first layer where the input is the photon counts ($\boldsymbol{v}^{(1)} = \boldsymbol{N}_T$), the hidden-input relationship is a Poisson restricted Boltzmann machine [32], described

below. For notation simplicity we omit the layer superscript.

$$P(N_{i,T}|\boldsymbol{h}) = Poiss(N_{i,T}|\exp(\sum_j h_j W_{ij} + b_i^V)T), \tag{4.4}$$

where $W \in \mathbb{R}^{n_V \times n_H}$ and $b^V \in \mathbb{R}^{n_V}$ are weights and biases of the model. Since the connectivity is local, for each column in W, which corresponds to a hidden unit, only a small set (e.g. 25) of the entries are non-zero. The hidden units collectively model the mean firing rate $\lambda_i = \exp(\sum_j h_j W_{ij} + c_i^V)$ on location i.

Conversely conditioning on the cumulative photon count $\boldsymbol{N}_T$, the hidden units become independent and their distribution is given by:

$$P(h_j = 1|\boldsymbol{N}_T) = Sigm(\sum_i N_{i,T} W_{ij} + b_j^H). \tag{4.5}$$

Inference on the deep belief network faces one critical issue, which is that the observations are evolving over time, i.e. we need to compute $P(h_j = 1|\boldsymbol{N}_t)$ for any $t \leq T$, instead of merely the highly-exposed 'image' at time T. This may be done by marginalizing out the unobserved counts $\Delta\boldsymbol{N} \triangleq \sum_{t'=t+1}^{T} \boldsymbol{X}_{t'}$:

$$P(h_j = 1|\boldsymbol{N}_t) = \sum_{\Delta\boldsymbol{N}} Sigm(\sum_i (N_{i,t} + \Delta N_i) W_{ij} + b_j^H) P(\Delta\boldsymbol{N}|\boldsymbol{N}_t) \tag{4.6}$$

$$\approx Sigm(\sum_i (N_{i,t} + (T-t)\mathbb{E}[\lambda_i|N_{i,t}]) W_{ij} + b_j^H), \tag{4.7}$$

where $\mathbb{E}[\lambda_i|N_{i,t}]$ is the estimated firing rate for location i. Using a Gamma prior $Gam(\mu_i t_0, t_0)$ on λ_i[1] we obtain that

$$P(h_j = 1|\boldsymbol{N}_t) \approx Sigm(\alpha(t) \sum_i W_{i,j} N_{i,t} + \beta_j(t)),$$

where $\alpha(t) \triangleq \frac{T+t_0}{t+t_0}$ and $\beta_j(t) \triangleq \frac{\tau(T-t)}{t+t_0} \sum_i W_{ij}\mu_i + b_j^H$ are two smooth scalar functions in t. Detailed derivations are in **Sec. A.2**.

Therefore, the log posterior ratio of the hidden units at the first layer is given by:

$$S_j^H(\boldsymbol{N}_t) \triangleq \log \frac{P(h_j = 1|\boldsymbol{N}_t)}{P(h_j = 0|\boldsymbol{N}_t)} \approx \alpha(t) \sum_i W_{i,j} N_{i,t} + \beta_j(t). \tag{4.8}$$

The log posterior ratio of the pooling unit m_k is:

$$S_k^M(\boldsymbol{N}_t) \triangleq \log \frac{P(m_k = 1|\boldsymbol{N}_t)}{P(m_k = 0|\boldsymbol{N}_t)} \quad = \underset{j \in G_k}{\mathcal{S}\max} \left(S_j^H(\boldsymbol{N}_t)\right) \approx \max_{j \in G_k} S_j^H(\boldsymbol{N}_t), \tag{4.9}$$

which is identical to the standard max pooling and the Maxout nonlinearity in deep networks [33], [34].

[1] We use a Gamma prior because it is the conjugate prior of the Poisson likelihood.

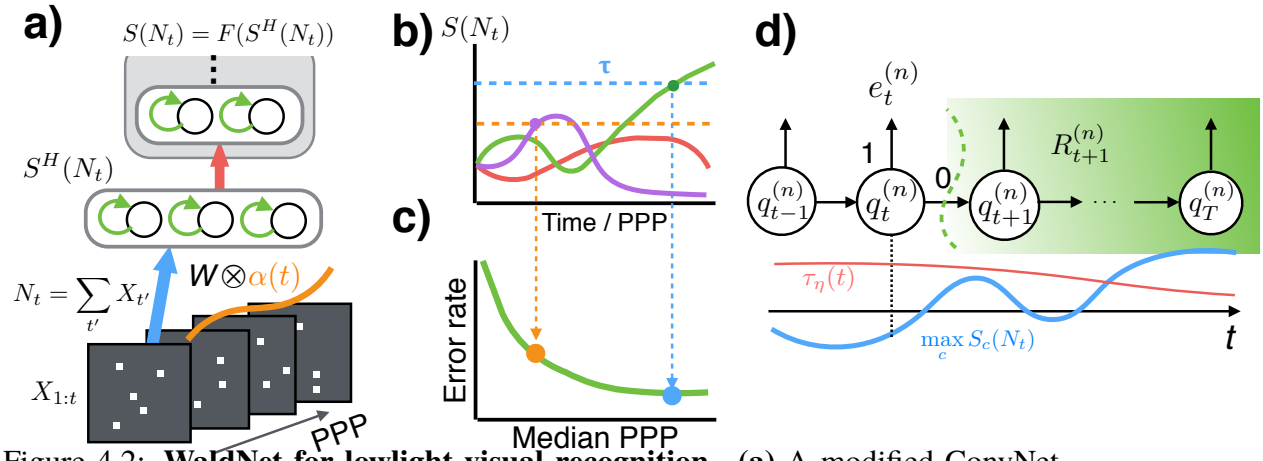

Figure 4.2: **WaldNet for lowlight visual recognition.** **(a)** A modified ConvNet for computing class posterior. The first layer is adapted (**Eq. 4.10**) to capture time-invariant features. From the cumulative photon counts $\boldsymbol{N}_t$ from duration $[0, t\Delta]$ (visualization in **Fig. 4.1**), WaldNet approximately computes hidden features $S^H(\boldsymbol{N}_t)$ that marginalize over unseen photons using weights $\boldsymbol{W}$ scaled by a time-varying scalar function $\alpha(t)$ (**Eq. 4.8**)). It then feeds the features into the remainder of the ConvNet F to compute log class posterior ratio $S(\boldsymbol{N}_t)$. **(b)** Deciding when to stop collecting photons. The class posteriors race to a common threshold to determine the category to predict. WaldNet stops photon collection as soon as one class crosses the threshold (**Eq. 4.3**). The example shows $S(\boldsymbol{N}_t)$ for three classes where the true class is green. Using a higher threshold (blue) yields a later but more accurate solution whereas a lower (orange) threshold is faster but risks misclassification. **(c)** The SAT curve (illustration only) produced by repeating (a-b) for multiple images and sweeping the threshold τ. **(d)** Learning time-varying threshold $\tau_\eta(t)$ (when class posterior learning (**Eq. 4.12**) is imperfect) to optimize Bayes risk with cost of error η (**Eq. 2.1**). The centipede network describes the recurrence relationship between risk $R_t^{(n)}$ starting from time t of example n and the risk $R_{t+1}^{(n)}$ starting from time $t+1$ (**Eq. 4.13**). $q_t^{(n)}$ is a gate (based on whether $S(\boldsymbol{N}_t)$ crosses threshold) that decides whether WaldNet stops at t with misclassification risk $e_t^{(n)}$ or continues collecting photons with risk $R_{t+1}^{(n)}$.

Discriminative training of WaldNet

Since the generative model may not be available in many practical applications, it may be more convenient to train a classifier that directly predicts the log posterior ratio $S(\boldsymbol{N}_t)$ and that shares the same computational structure as the inference procedure of the generative model. Fortunately the inference procedure bears striking similarity to a ConvNet, so that powerful deep learning tools (e.g. provided by the MatConvNet toolbox [35]) may be applied. Now we present the discriminative reasoning.

Inference procedure

Recall from the previous section that the inference procedure of WaldNet is an adjusted version of the standard ConvNet. In ConvNet, the input is an image $\boldsymbol{N}_T$ obtained from a fixed observation time T. ConvNet contains multiple layers of computations that may be viewed as a nesting of two transformations: (1) the first hidden layer $S^H(\boldsymbol{N}_T) = \mathbf{W}\boldsymbol{N}_T + \mathbf{b}^H$ that maps the input to a feature vector, and (2) the remaining layers $S(\boldsymbol{N}_T) = F(S^H_T)$ that map the features S^H to the log class posterior probabilities $S(\boldsymbol{N}_T)$. $\mathbf{W} \in \mathbb{R}^{D\times n_H}$ is a weight vector and $\mathbf{b}^H \in \mathbb{R}^{n_H}$ is a bias vector.

WaldNet differs from a ConvNet in two aspects. (1) The input $\boldsymbol{N}_t$ to a WaldNet is a *time-series* that includes the cumulative photon counts up to a moving horizon t, and the output $S(\boldsymbol{N}_t)$ is also a time-series, which encodes the log class posterior probabilities over time. (2) The first-layer features in WaldNet are computed differently *depending on the exposure time* t. The weights and biases of the transformation in S^H are adjusted smoothly over time using $\alpha(t) \in \mathbb{R}$ and $\boldsymbol{\beta}(t) \in \mathbb{R}^{n_H}$ (see **Eq. 4.8** and **Eq. 4.9**):

$$S^H(\boldsymbol{N}_t) = \alpha(t)\boldsymbol{W}\boldsymbol{N}_t + \boldsymbol{\beta}(t), \tag{4.10}$$

while the rest of the computations stays the same: $S(\boldsymbol{N}_t) = F(S^H(\boldsymbol{N}_t))$.

> The main intuition of our approach is that the stochasticity in photon arrivals is addressed with an exposure-time specific transformation S^H, and the intra- and inter- class variation is captured with an exposure-time invariant transformation F. The revised network has nearly the same number of parameters as a conventional ConvNet, but has the capacity to process inputs at different exposure times. The adaptation is critical for performance, as will be seen by comparison with simple rate-based methods in **Sec. 4.4**.

> Why do we single out the first layer features $S^H(\boldsymbol{N}_t)$ for adjustment? In theory features at any layer would do but it is more convenient at the first layer. This is because the adjustment procedure uses mean-field approximations and this (1) becomes increasingly less accurate as the feature computation becomes more nonlinear, and (2) requires computing the posterior mean of the feature, which may not have a handy closed form.

Training strategy

Recall that our goal is to train WaldNet to optimize the Bayes risk [36] (**Eq. 2.1**). In

scotopic vision the Bayes risk R is formulated as

$$R \triangleq \mathbb{E}[t] + \eta \mathbb{E}[C \neq \hat{C}_t], \tag{4.11}$$

where $\mathbb{E}$[t] is the expected photon count required for classification, $\mathbb{E}[C \neq \hat{C}_t]$ is the error rate, and η describes the user's cost of error versus time. WaldNet asymptotically optimizes the Bayes risk provided that it can faithfully capture the log class posterior ratio $S(\boldsymbol{N}_t)$, and selects the correct threshold τ (**Eq. 2.3**). Sweeping η allows WaldNet to traverses the optimal SAT (**Fig. 4.2c**).

Our strategy is to separate training into two steps with distinct objectives: step one trains a WaldNet to approximate $S(\boldsymbol{N}_t)$, and step two picks the optimal threshold according to η to minimize the Bayes risk.

Step one: posterior learning
Given a lowlight dataset $\{\boldsymbol{N}_t^{(n)}, C^{(n)}\}_{n,t}$ where n indexes training examples and t indexes exposure time, we train the WaldNet to minimize:

$$-\sum_{n,t} \log P(C = C^{(n)} | \boldsymbol{N}_t^{(n)}, \mathcal{W}) + reg(\mathcal{W}), \tag{4.12}$$

where $\mathcal{W}$ collectively denote all the parameters in the WaldNet, and $reg(\mathcal{W})$ denotes $L2$ weight-decay on the filters. When a lowlight dataset is not available we simulate the dataset from intensity images according to the noise model in **Eq. 4.1**, where the exposure times are sampled uniformly on a logarithmic scale (see **Sec. 4.4**).

Step two: threshold tuning
After step one, if WaldNet captures the log class posterior ratios $S(\boldsymbol{N}_t)$, we can simply optimize a scalar threshold τ_η for each tradeoff parameter η. In practice, we may opt for a time-varying threshold $\tau_\eta(t)$ as step one may not be perfect.

> For instance, consider an adapted ConvNet that perfectly captures the class posterior. Ignoring the regularizer (right term of **Eq. 4.12**), we can scale up the weights and biases of the last layer (softmax) by an arbitrary amount without affecting the error rate, which scales the negative log likelihood (left term in **Eq. 4.12**) by a similar amount, leading to a better objective value. The magnitude of the weights are thus determined by the regularizer and may be off by a scaling factor. We therefore need to properly rescale the class posterior at every exposure time before comparing to a constant threshold, which is equivalent to using a time-varying threshold $\tau_\eta(t)$ on the raw predictions.

To learn the time-varying threshold $\tau_\eta(t)$, we need to formulate the Bayes risk objective as a function of $\tau_\eta(t)$. Let $\{\boldsymbol{N}_t^{(n)}\}_{t=1}^T$ be a sequence of lowlight images that are increasing in exposure time and generated from the n-th intensity image. Denote $q_t^{(n)} \triangleq \mathbb{I}[\max_c S_c(\boldsymbol{N}_t) > \tau_\eta(t)]$ the event that the posterior crosses decision threshold at time t, and $e_t^{(n)}$ the event that the class prediction at t is wrong. Let $R_t^{(n)}$ denote the Bayes risk of the sequence (indexed by n of the high-quality image $X^{(n)}$) incurred from time t onwards. $R_t^{(n)}$ may be computed recursively:

$$R_t^{(n)} = \Delta + \eta \left(q_t^{(n)} e_t^{(n)} + (1 - q_t^{(n)}) R_{t+1}^{(n)} \right), \tag{4.13}$$

where the first term is the cost of collecting photons during time interval $((t-1)\Delta t\Delta]$, the second term is the expected cost of committing to a decision that is wrong, and the last term is the expected cost of deferring the decision till more photons are collected.

The Bayes risk is obtained from averaging multiple photon count sequences, i.e. $R = \mathbb{E}[R_0^{(n)}]$. $q_t^{(n)}$ is non-differentiable with respect to the threshold $\tau_\eta(t)$, leading to difficulties in optimizing R. Instead, we approximate $q_t^{(n)}$ with a Sigmoid function,

$$q_t^{(n)}(\tau_\eta(t)) \approx Sigm\left(\frac{1}{\sigma_{temp}} (\max_c S_c(\boldsymbol{N}_t) - \tau_\eta(t)) \right), \tag{4.14}$$

where $Sigm(x) \triangleq 1/(1+\exp(-x))$, and anneal the temperature σ_{temp} of the Sigmoid over the course of training [37] (see **Sec. 4.4**).

> Even though we assume a certain form for the log class posterior ratio $S(\boldsymbol{X}_{1:t})$, this threshold learning procedure is very general and works for any $S(\boldsymbol{X}_{1:t})$. In particular, it may be used for learning SPRT procedures when the underlying probabilistic distribution is not i.i.d. in time.

4.4 Experiments

Exposure time versus signal

Our experiments use PPP interchangeably with exposure time t for performance measurement, since PPP directly relates to the number of bits of signal in each pixel (**Eq. 4.2**). In practice an application may be more concerned with exposure time. Thus it is helpful to relate exposure time, PPP and the bits of signal. **Table 4.1** describes this relationship for different illuminance levels. Derivations are in the Appendix **Sec. A.2**.

Scene	Illuminance E_v (LUX)	exposure time t (s) 1/500	1/128	1/8	1	8	60
Moonless	10^{-3}					1.5	3
Full moon	1	0.5	1.5	3.5	5	6.5	8
Office	250	4.5	5.5	7.5	9	10.5	12
Overcast	10^3	5.5	6.5	8.5	10	11.5	13
Bright sun	10^5	9	10	12	13.5	15	16.5

Table 4.1: (Approximate) number of bits of signal per pixel under different illuminance levels. See Appendix for full derivation. For instance, in an office scene it takes $1/8$ seconds to obtains a 7.5-bit image. Under full moon, the same high-quality image and the same sensor needs > 8 seconds to capture.

Baseline Models

We compare WaldNet against the following baselines:

Ensemble. We construct the ensemble (**Sec. 4.3**) using "specialist" models. Each specialist is a ConvNet with the same model dimensions (number of layers, number of hidden units of each layer, nonlinearity, etc) as the WaldNet, but is trained using only cumulative photon counts at a single PPP. We use four specialists with PPPs from $\{.22, 2.2, 22, 220\}$ respectively. To test cumulative counts $\boldsymbol{N}_{PPP'}$ with a PPP that is not on the training set, we rescale $\boldsymbol{N}_{PPP'}$ to have the same PPP as the specialist with the closest PPP. As the number of specialists grows, the ensemble approaches the best achievable SAT for WaldNet.

Photopic classifier. To justify the necessity of modeling photon count statistics in lowlight, we introduce another intuitive classifier. The classifier is a ConvNet trained on 'images' $\boldsymbol{N}_T$ from normal lighting conditions, and applied to properly rescaled cumulative counts $\boldsymbol{N}_t$ for $t \leq T$. We choose the specialist with PPP= 220 as the photopic classifier as it achieves the same accuracy as a network trained with 8-bit images.

Rate classifier. To test the significance of the uncertainty information carried by the cumulative counts, we train a classifier directly on the rate estimates without weight adaptation. Formally, the hidden unit on layer one is $S^H(\boldsymbol{N}_t) \approx \boldsymbol{W}\boldsymbol{N}_t/t + \boldsymbol{b}_j^H$. Note the similarity with our approximation used in **Eq. 4.8**.

We assume that all models have an internal clock, which enables the model to estimate the expected PPP under the constant illuminance assumption. When the illuminance changes, the model may rely on an independent external measure or the cumulative count itself to adjust PPP.

We consider two standard datasets: MNIST [17] and CIFAR10 [16]. We simulate

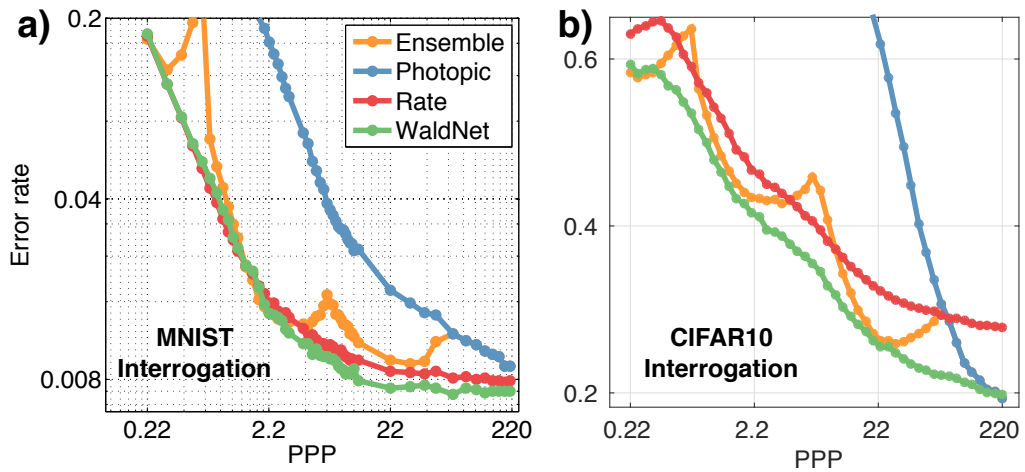

Figure 4.3: **Interrogation performance comparison**. Error rate plotted against the interrogation PPP for (**a**) MNIST and (**b**) CIFAR10. Each dot is computed from classifying $10k$ test examples with a fixed PPP.

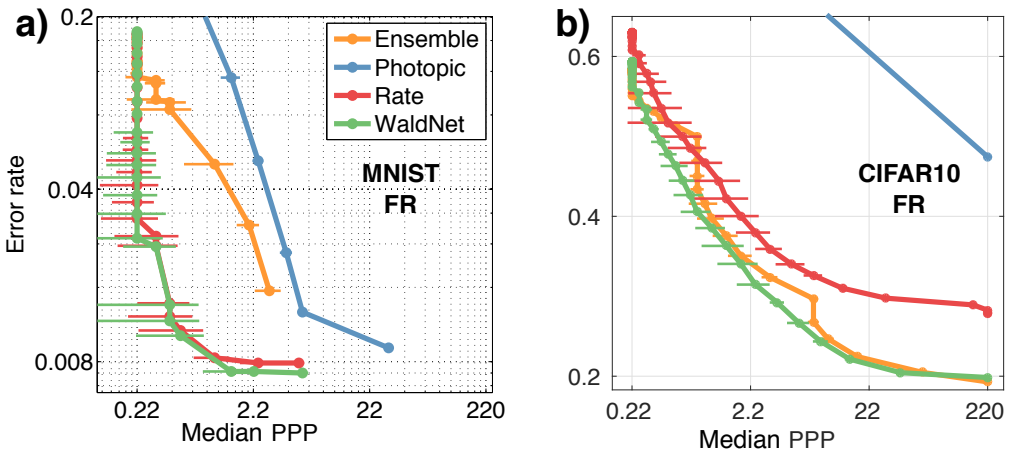

Figure 4.4: **Free response performance comparison**. Error rate plotted against *median* PPP for (**a**) MNIST and (**b**) CIFAR10. 1 bootstrap *ste* is shown for both the median PPP and error rate, the latter is too small to be visible.

lowlight image sequences using **Eq. 4.1**. MNIST contains gray-scaled 28×28 images of 10 hand-written digits. CIFAR10 contains 32×32 color images of 10 visual categories. The details of model architectures and training procedure are found in the Appendix **Sec. A.2**.

Results

The SAT curves in the INT regime are shown in **Fig. 4.3a** and b. Median PPP versus accuracy tradeoffs for all models in the FR regime are shown in **Fig. 4.4a** for MNIST and **Fig. 4.4b** for CIFAR10. All models use constant thresholds for

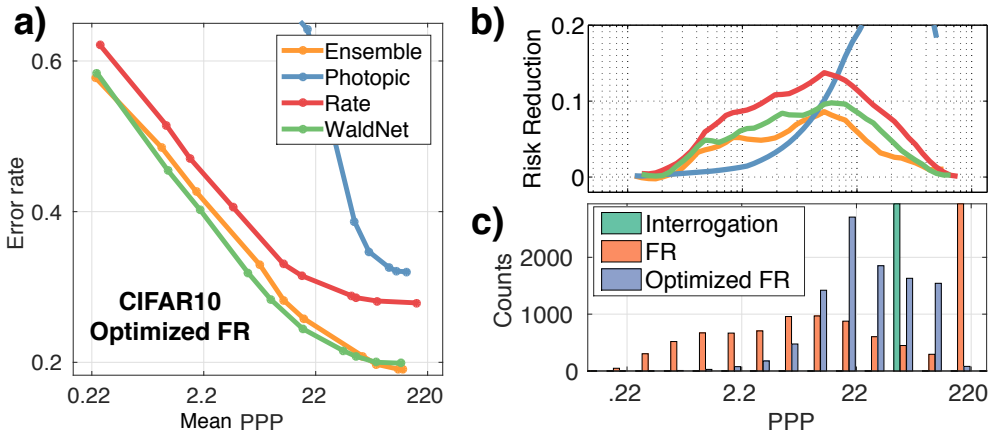

Figure 4.5: **Effect of threshold learning (Sec. 4.3)**. (**a**) Error rate against the *average* PPP for CIFAR10 using a network with optimized time-varying threshold $\tau_\eta(t)$. 1 bootstrapped *ste* is shown but not visible. (**b**) Each curve shows the Bayes risk reducation after optimization (**Sec. 4.3**, step two) per *average* PPP. (**c**) Response time (PPP) histograms under the interrogation, FR (before optimization), and FR (after optimization) of a WaldNet that achieves 22% error on CIFAR10.

producing the tradeoff curves. In **Fig. 4.5a** are average PPP versus accuracy curves when the models use the optimized dynamic thresholds (**Sec. 4.3**, step two).

Model comparisons

Overall, WaldNet performs well under lowlight. It only requires < 1 PPP to stay within 0.1% (absolute) degradation in accuracy on MNIST and around 20 PPP to stay within 1% degradation on CIFAR10, even though recognition at such light levels (**Fig. 4.1**) may prove difficult for humans.

The ensemble was formed using specialists at logarithmically-spaced exposure times, thus its curve is discontinuous in the INT regime (**Fig. 4.3**). The peaks delineate transitions between specialists. The ensemble's performance at the specialized light levels $[.22, 2.2, 22, 220]$ also provides a proxy for the performance upper bound by ConvNets of the same architecture (apart from overfitting and convergence issues during learning). Using this proxy we see that even though WaldNet uses $1/4$ the parameters of the ensemble, it stays close to the performance upper bound. In FR regime, the ensemble is outperformed by WaldNet on MNIST (due to overfitting) and on par on CIFAR10 for lowlight conditions (< 22 PPP). This showcases WaldNet's ability to handle photon counts at multiple PPPs without requiring explicit parameters (as it is the case for the ensemble).

The photopic classifier retrofitted to lowlight applications does not work well in

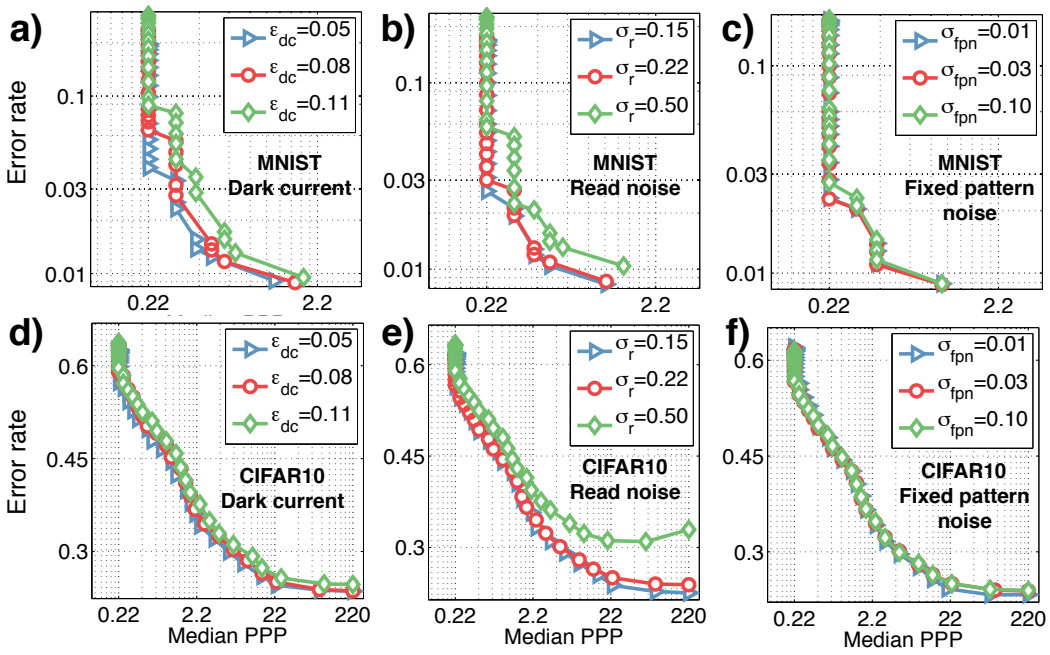

Figure 4.6: **Effect of sensor noise on WaldNet**. The rows correspond to datasets MNIST and CIFAR10, and the columns correspond to parameters of noise sources, which are the dark current ϵ_{dc}, the standard deviation of additive read noise σ_r, and the standard deviation of multiplicative fixed pattern noise σ_{fpn}. The baseline has $\epsilon_{dc} = 3\%$ and $\sigma_r = \sigma_{fpn} = 0$ for MNIST, and $\epsilon_{dc} = 5\%$, $\sigma_r = 0.22$ and $\sigma_{fpn} = 0.03$ for CIFAR10.

either dataset, which showcases the necessity of WaldNet as well as training with scotopic input. On MNIST, the photopic classifier also underperforms WaldNet in highlight regimes. This is because MNIST is rather easy to overfit, and training with lowlight inputs provides a form of regularization.

The rate classifier differs from WaldNet only in how the first layer feature is computed, thus the better performance of WaldNet in CIFAR10 is due solely to the WaldNet's time-adapted features (**Eq. 4.8**).

Effect of threshold learning

With constant thresholds (**Fig. 4.4**) WaldNet significantly outperforms the photopic classifier. As the latter has never seen any lowlight inputs, its assessment of the log posterior ratio is ill-suited to SPRT. Using learned dynamic thresholds (step two of **Sec. 4.3**) we see consistent improvement on the *average* PPP required for given error rate across all models (**Fig. 4.5b**), with more benefit for the photopic classifier. **Fig. 4.5c** examines the PPP histograms on CIFAR10 with constant (FR) versus dynamic threshold (optimized FR). We see with constant thresholds many

decisions are made at the PPP cutoff of 220, so the median and the mean are vastly different. Learning dynamic thresholds reduce the variance of the PPP but make the median longer. This is ok because the Bayes risk objective (**Eq. 2.1**) concerns the average PPP, not the median. Clearly which threshold to use depends on whether the median or the mean is more important to the application.

Effect of INT versus FR

Cross referencing **Fig. 4.3** and **Fig. 4.4** reveals that FR with constant thresholds often brings 3x reduction in median photon counts. Dynamic thresholds also produce faster *average* and *median* responses. This is consistent with our theoretical result in **Theorem. 1**.

Sensitivity to sensor noise

Finally, we inspect how the network's performance is affected by sensor noise. For MNIST and CIFAR10, we take WaldNet and vary independently the dark current, the read noise and the fixed pattern noise (**Fig. 4.6**).

First, the effect of dark current and fixed pattern noise is minimal. Even an 11% dark current (i.e. photon emission rate of the darkest pixel is 10% of that of the brightest pixel) merely doubles the exposure time with little loss in accuracy. The multiplicative fixed pattern noise does not affect performance because WaldNet in general makes use of very few photons. Second, current industry standard of read noise ($\sigma_r = 22\%$ [9]) guarantees no performance loss. Lastly, the fact that $\sigma_r = 50\%$ hurts performance suggests that single-photon resolution is vital for scotopic vision (**Fig. 4.6b,e**).

4.5 Chapter summary

We proposed to study the important yet relatively unexplored problem of scotopic visual recognition. Scotopic vision is vision starved for photons. This happens when available light is low, and image capture time is longer than computation time. In this regime vision computations should start as soon as the shutter is opened, and algorithms should be designed to process photons as soon as they hit the photoreceptors. While visual recognition from limited evidence has been studied [38], to our knowledge, our study is the first to explore the exposure time versus accuracy trade-off of visual classification, which is essential in scotopic vision.

We proposed WaldNet, a model that combines photon arrival events over time to form a coherent probabilistic interpretation, and make a decision as soon as sufficient

evidence has been collected. The proposed algorithm may be implemented by a deep feed-forward network similar to a convolutional network. Despite the similarity of architectures, we see clear advantages of approaches developed specifically for the scotopic environment. An experimental comparison between WaldNet and models of the conventional kind, such as photopic approaches retrofitted to lowlight images and ensemble-based approaches agnostic of lowlight image statistics, shows large performance differences, both in terms of model parsimony and response time (measured by the number of photons required for decision at desired accuracy). Finally, despite relying only on few photons for decisions, WaldNet is minimally affected by camera noises, making it an ideal model to be integrated with the recently-developed lowlight sensors.

References

[1] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Submitted to IEEE Trans. on Pattern Anal. and Machine Intell.*, 2013.

[2] P. A. Morris, R. S. Aspden, J. E. Bell, R. W. Boyd, and M. J. Padgett, "Imaging with a small number of photons," *Nature Communications*, vol. 6, 2015.

[3] C. Ferree and G. Rand, "Intensity of light and speed of vision: I.," *Journal of Experimental Psychology*, vol. 12, no. 5, p. 363, 1929.

[4] E. D. Dickmanns, *Dynamic vision for perception and control of motion*. Springer Science & Business Media, 2007.

[5] S. Thorpe, D. Fize, C. Marlot, *et al.*, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[6] D. J. Stephens and V. J. Allan, "Light microscopy techniques for live cell imaging," *Science*, vol. 300, no. 5616, pp. 82–86, 2003.

[7] E. Hall and D. Brenner, "Cancer risks from diagnostic radiology," *Cancer*, vol. 81, no. 965, 2014.

[8] L. Sbaiz, F. Yang, E. Charbon, S. Süsstrunk, and M. Vetterli, "The gigavision camera," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, pp. 1093–1096.

[9] E. Fossum, "The quanta image sensor (qis): Concepts and challenges," in *Imaging Systems and Applications*, Optical Society of America, 2011, JTuE1.

[10] F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.

[11] H. Barlow, "A method of determining the overall quantum efficiency of visual discriminations," *The Journal of Physiology*, vol. 160, no. 1, pp. 155–168, 1962.

[12] T. Delbrück and C. Mead, "Analog vlsi phototransduction," *Signal*, vol. 10, no. 3, p. 10, 1994.

[13] J. I. Gold and M. N. Shadlen, "Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward," *Neuron*, vol. 36, no. 2, pp. 299–308, Oct. 2002.

[14] B. Chen, V. Navalpakkam, and P. Perona, "Predicting response time and error rates in visual search," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

[15] M. N. Wernick and G. M. Morris, "Image classification at low light levels," *Journal of the Optical Society of America A*, vol. 3, no. 12, pp. 2179–2187, 1986.

[16] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*, 2009.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, IEEE, vol. 1, 2001, pp. I–511.

[19] P. Moreels, M. Maire, and P. Perona, "Recognition by probabilistic hypothesis construction," in *Computer Vision-ECCV 2004*, Springer, 2004, pp. 55–68.

[20] J. Matas and O. Chum, "Randomized ransac with sequential probability ratio test," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, IEEE, vol. 2, 2005, pp. 1727–1732.

[21] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman, "Automatic estimation and removal of noise from a single image," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 2, pp. 299–314, 2008.

[22] E. R. Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors," *Electron Devices Society, IEEE Journal of the*, vol. 1, no. 9, pp. 166–174, 2013.

[23] G. E. Healey and R. Kondepudy, "Radiometric ccd camera calibration and noise estimation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 3, pp. 267–276, 1994.

[24] L. Anzagira and E. R. Fossum, "Color filter array patterns for small-pixel image sensors with substantial cross talk," *Journal of the Optical Society of America A*, vol. 32, no. 1, pp. 28–34, 2015.

[25] C. W. Baum and V. V. Veeravalli, "A sequential procedure for multihypothesis testing," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, 1994.

[26] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests. ii. accurate asymptotic expansions for the expected sample size," *Information Theory, IEEE Transactions on*, vol. 46, no. 4, pp. 1366–1383, 2000.

[27] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks.," *Psychological Review*, vol. 113, no. 4, p. 700, 2006.

[28] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, *et al.*, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, p. 153, 2007.

[29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[30] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009, pp. 609–616.

[31] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[32] R. Salakhutdinov and G. Hinton, "Semantic hashing," *International Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 969–978, 2009.

[33] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[34] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *ArXiv preprint arXiv:1302.4389*, 2013.

[35] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, ACM, 2015, pp. 689–692.

[36] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[37] H. Mobahi and J. W. Fisher III, "On the link between gaussian homotopy continuation and convex envelopes," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Springer, 2015, pp. 43–56.

[38] S. M. Crouzet, H. Kirchner, and S. J. Thorpe, “Fast saccades toward faces: Face detection in just 100 ms,” *Journal of Vision*, vol. 10, no. 4, p. 16, 2010.

[1] B. Chen and P. Perona, “Scotopic visual recognition,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015, pp. 8–11.

Chapter 5

VISUAL DISCRIMINATION WITH UNKNOWN STIMULUS ONSET

Sequential Reasoning with a Nonstationary Probabilistic Model

Our last project is a psychophysics study of visual discrimination with uncertain stimulus onset. Unlike the previous problems, in this problem a probabilistic model is given, but the model is not stationary over time.

5.1 Motivation

An organism's survival is critically dependent on its ability to detect change (e.g. the sound/sight of something moving in the distance), and classify its nature (e.g. a predator, prey, or meaningless clutter). In ecological conditions, change detection and object classification frequently co-occur: approaching animals need to be detected and classified as friend or foe. Despite the ecological significance of considering detection and classification jointly, the two tasks are typically studied in isolation. Consequently, it remains unknown how humans jointly perform classification and detection, and whether and how humans trade off speed and accuracy.

Psychologists have studied the phenomenology of visual discrimination as well as computational approaches [1]–[3]. We have reviewed that the optimal model for trading off speed and accuracy is the sequential probability ratio test (SPRT) [4]. When the discrimination is between two simple templates, the diffuse-to-bound process [5] is also optimal. These discrimination models require knowing when change happens, i.e. when to start accumulating evidence, which is not a realistic hypothesis in most ecological conditions.

The phenomenology of change point detection is relatively less explored. Earlier studies examine *whether* change occurred [6]–[9], and, more recently, *when* it occurred [10], [11]. The optimal model for minimizing detection error and reaction time [12] dates back to the cumulative sum control chart (CUSUM) [13], [14], which utilizes a diffuse-to-bound mechanism with only one absorbing boundary. When the change could bring the world into one of multiple states, a network of diffusions is required [15] to integrate changes attributable to different categories optimally. Despite addressing the uncertainty in change onset, these models do not consider the question of classification.

Contributions

(1) We study the the **joint** detection and classification task (the 'dual task' for brevity). Our experiment is a variant of random dot motion discrimination [2] where the motion is completely incoherent at first. After a random delay it becomes coherent in one of two directions. The subject is asked to both detect change and classify the coherent motion. We manipulated the motion directions to control the relative difficulty of detection and classification.

(2) We developed three **computational models** for the dual task. The first model 'Classifies and then Detects' (CD), which is optimal [16]. CD applies SPRT on the probabilistic model of both classification and detection. The second and third models are computationally simpler and sub-optimal, where they apply SPRT separately on the detection and the classification problem. The two models differ in the temporal order in which the SPRT modules are executed. Model two conducts 'Detection and Classification in Parallel' (DCP), while the third model conducts 'Detection and Classification in Series' (DCS).

(3) We **test human subjects** on the dual task as well as a pure detection task. Fitting the parameters of our models to data collected from both tasks reveals that the only model that is consistent with human SAT behavior is the **conceptually simple but sub-optimal** DCS model. Primates have been found to be near-optimal in detection and classification [11], [17]–[19] and our findings deviate from this pattern.

(4) To fit our models on random-dot motion patterns, we develop a simple model of early vision [20], [21] based on **quantized** sensory input, which are action potentials from motion-tuning neurons in area MT [22]. This model is parsimonious and versatile: with one free parameter it simulates sensory inputs for detection tasks and dual tasks with arbitrary coherent strengths and motion directions. This generalization ability is an improvement over other decision models of random dot motion discrimination [3], [23], [24], which typically are independently parameterized across tasks and only generalize across the level of coherence of motion stimuli.

5.2 Framework for visual discrimination with unknown onset

Chapter-specific notations

Formally in the dual task, the world exists in one of three states at any given time bin t: $C_t \in \{0, 1, 2\}$, where time bin t represents the duration $((t-1)\Delta, t\Delta]$, where 0 is the initial state (e.g. incoherent motion), 1 and 2 are two post-change states to be

distinguished (e.g. coherent motion along one of two directions). The world always starts from $C_0 = 0$ and changes to either class 1 or class 2 at a random time t_δ . The change occurs *only once*. The observer has information regarding the distribution of the change time t_δ, but not the actual value of t_δ. The goal is to infer the stimulus category $C \in \{1, 2\}$ as quickly as possible, but not earlier than t_δ, in which case the response is considered a false detection error. **Fig. 5.1a** illustrates the setup for the dual task in the context of random dot motion discrimination (see **Sec. 5.3**).

Models

Our three models (CD, DCP, and DCS) vary in optimality and simplicity. CD is optimal. The initial incoherent motion and the two coherent motions are modeled as three separate stimulus categories, and the dual task is reduced to a multi-category classification task, which may be solved optimally [16]. DCS and DCP are computationally simpler and sub-optimal. Both use a detector to identify any kind of coherent motion, and a classifier to distinguish between the two motion directions. In DCP, the detector and the classifier operate **simultaneously**, and as soon as the detector reveals a change, the classifier is consulted to reveal the nature of the change. In DCS, the coherent motion detector **triggers** the integration time for a classical diffuse-to-bounds classifier which eventually reaches a decision.

Our models of the dual task assume optimal evidence accumulation from input sensors [21]: all models have access to the following two statistics computed according to Bayesian inference. The first statistic is the log posterior ratio between any pair of classes (derived in **Eq. 5.10**):

$$S_t^{i,j} \triangleq \log \frac{P(C_t = i|\boldsymbol{X}_{1:t})}{P(C_t = j|\boldsymbol{X}_{1:t})}, \tag{5.1}$$

where $S_t^{i,j}$ is the log posterior ratio between class i and j ($i, j \in \{0, 1, 2\}$) given evidence $\boldsymbol{X}_{1:t}$ collected up to time t. We overload this notation to represent ratios between sets of classes. For example, $S_t^{1,\bar{1}}$ means the log posterior ratio of class 1 versus 'not 1', which contains class 2 and class 0.

The other important statistics is the log posterior ratio between the coherent motion classes (class 1 and 2) assuming that *the change has occurred* at time t (derived in **Eq. 5.13**):

$$R_{t_\delta,t} \triangleq \log \frac{P(C_t = 1|\boldsymbol{X}_{t_\delta:t})}{P(C_t = 2|\boldsymbol{X}_{t_\delta:t})}, \tag{5.2}$$

where the log posterior ratio is conditioned on observations in the time interval $[t_\delta, t]$, e.g. from the time of change t_δ and a later time t.

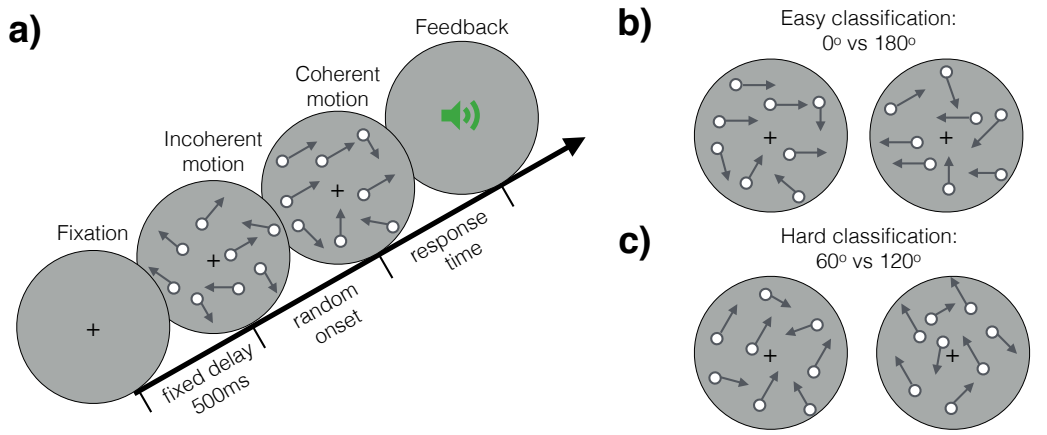

Figure 5.1: **Random dot motion discrimination with unknown stimulus onset**. (**a**) Stimulus setup. A trial begins with a central fixation cross. After $500ms$ a display of dots moving incoherently in all directions is displayed. After a random delay t_δ, a fraction z of the dots start moving coherently along one of two directions $\{\theta_1, \theta_2\}$. As quickly as possible the subject presses on a button to indicate the direction of motion. The trial ends with auditory feedback. The direction of coherent motion controls the relative difficulty between classification and detection. (**b**) Stimulus for coherent motion 0° and 180° (classification is easier than detection). (**c**) Stimulus for coherent motion 60° and 120° (detection easier than classification).

Both log posterior ratio statistics may be computed directly from the firing patterns of motion-tuning neurons in MT, to be discussed in the MT front-end section (e.g. **Eq. 5.10** and **Eq. 5.13**). Based on these statistics, we present three plausible models for the dual task.

Classify then Detect (CD)

The first system (**Fig. 5.2a,b**) is based on the posterior probability ratio of classes 1 and 2. The system employs two accumulators $S_t^{c,\bar{c}}$, one for each class $c \in \{1, 2\}$, that race to reach a threshold τ_{dis}. The class of the winner is the predicted class $\hat{C}$. Since in our tasks the two classes are completely symmetrical the same threshold τ_{dis} is set for both accumulators. Distinct thresholds may be necessary in asymmetric scenarios (e.g. one class is more frequent than the other). Let t_d denote the time of decision ($t_d - t_\delta$ is the reaction time). The CD procedure is:

$$\begin{aligned} t_c &= \text{first time } t \text{ that } S_t^{c,\bar{c}} > \tau_{dis}, \qquad c \in \{1, 2\} \\ \hat{C} &= \arg\min_{c \in \{1,2\}} t_c, \qquad t_d = t_{\hat{C}}. \end{aligned} \tag{5.3}$$

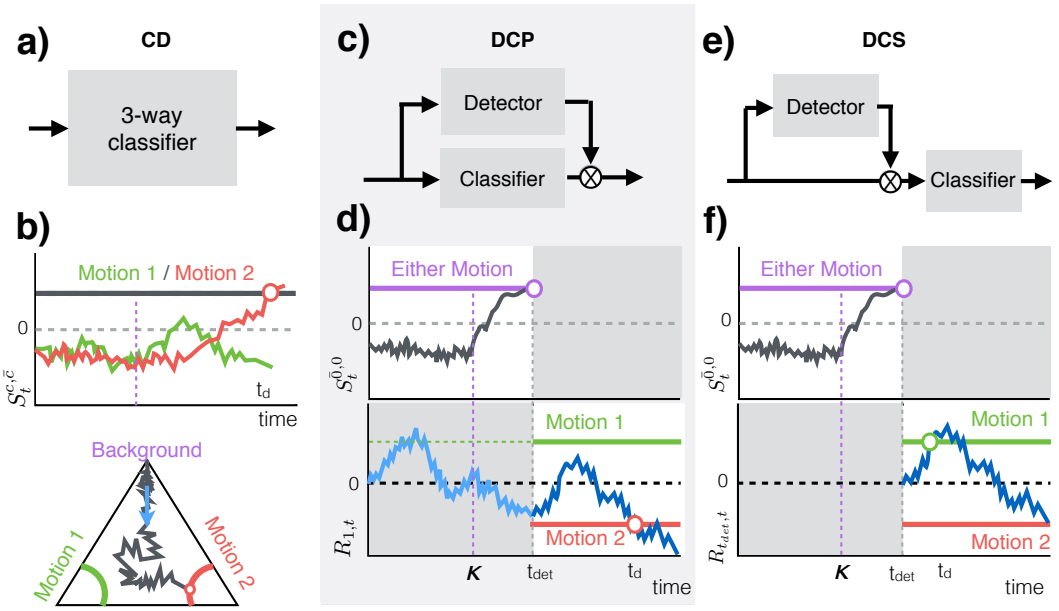

Figure 5.2: **Three models of joint detection and discrimination**. (**a**) **CD** – The world is assumed to be in one of three states: initial state (incoherent motion) and two post-change states (motion 1 and motion 2). The posterior probability of each state is computed. (**b**) (Top) In CD, log posterior ratios $S_t^{1,\bar{1}}$ and $S_t^{2,\bar{2}}$ (**Eq. 5.1**) race to a common discrimination threshold. (Bottom) An equivalent depiction showing the trajectory of the posterior over time (time direction indicated by the blue arrow) visualized in the probability simplex. When the posterior reaches one of the two lower corners the system declares motion 1 (left corner) or 2 (right corner). (**c**) **DCP** – A detector for coherent motion and a classifier of motion direction are computed in parallel. (**d**) The detector computes the log posterior ratio $S_t^{\bar{0},0}$ of any coherent motion vs. incoherent motion. The classifier computes the log posterior ratio $R_{1,t}$ of motion 1 vs. motion 2, which races towards a pair of thresholds (upper for motion 1, lower for motion 2). Until the detector fires at t_{det}, the classifier cannot fire (despite crossing dashed green threshold). After t_{det}, the classifier carries over signals prior to t_{det}. (**e**) DCS employs a detector and a classifier in series. (**f**) The classifier starts only after the detector fires and does not retain information prior to t_{det}. This lossy integration causes DCS to make a different (wrong in this example) decision than DCP.

This procedure is a Bayesian version of the multi-class CUSUM procedure [13] and proven optimal by [16]. Here optimality means that given a requirement on the false detection rate and the misclassification rate, the procedure above achieves the shortest response time on average.

Detection and Classification in Parallel (DCP)

The second model (**Fig. 5.2c,d**) separately and simultaneously performs detection

and discrimination. A detector performs a one-sided test on $S_t^{\bar{0},0}$, the log posterior ratio of 'coherent motion' (state 1 and 2) against state 0 of incoherent motion, to detect whether any coherent motion is present. Meanwhile, running in the background is a classifier that is concerned only with distinguishing between the two coherent motion classes $R_{1,t}$. The classifier is suppressed from firing until the detector fires at time t_{det}.

The decision process is parameterized by the threshold τ_{det} for detection, and the threshold τ_{dis} for classification. Again the discrimination threshold τ_{dis} is shared between classes for simplicity.

$$\begin{aligned} t_{det} &= \text{first time } t \text{ that } S_t^{\bar{0},0} > \tau_{det} \\ t_c &= \text{first time } t \geq t_{det} \text{ that } R_{1,t} > \tau_{dis}, c \in \{1,2\} \\ \hat{C} &= \arg\min_{c\in\{1,2\}} t_c, \qquad t_d = t_{\hat{C}} \end{aligned} \tag{5.4}$$

Here the detector and classifier run in parallel, and the detector functions as a gate that guards the classifier against fluctuations. Both the detector and the classifier are *lossless* in information integration, but the classifier is used sub-optimally since the information accumulated prior to stimulus onset is invalid.

> The DCP model may seem redundant as it is not optimal. It is included because of reverse compatibility and model complexity. First, DCP contains specialized and optimal components for detection and classification, respectively. By selecting the corresponding component DCP can solve pure detection or pure discrimination tasks. Second, DCP contains two independent decision thresholds, making it as complex as DCS (next subsection). Therefore any performance discrepancy between the two is directly attributable to model biases, not complexity.

Detection and Classification in Series (DCS)

In the third model, the last we consider, model detection and classification proceed in succession (**Fig. 5.2e,f**). After the detector identifies a coherent motion at time t_{det}, the classifier comes online assuming that the change has already happened ($t_\delta \leq t_{det}$). This assumption reduces the problem to pure classification starting at time t_{det}, which may be solved by the classical sequential probability test (SPRT [4]).

$$
\begin{aligned}
t_{det} &= \text{first time } t \text{ that } S_t^{\bar{0},0} > \tau_{det} \\
t_c &= \text{first time } t \text{ that } R_{t_{det},t} > \tau_{dis}, c \in \{1,2\} \\
\hat{C} &= \arg\min_{c\in\{1,2\}} t_c, \qquad t_d = t_{\hat{C}}
\end{aligned} \tag{5.5}
$$

DCS essentially concatenates the optimal detector (CUSUM) and the optimal classifier (SPRT) in time. It is *lossy* and potentially slower because the classifier does not consider any evidence before the detector fires. However it also provides *modularity*, as it completely separates the detection problem from discrimination. Mathematically, the subtle difference between DCS (Equation set 5.5) and DCP (Equation set 5.4) is the lossy evidence accumulation by the classifier. DCS discards all observations prior to detector firing, while DCP maintains them. Therefore, comparing DCP and DCP allows us to understand whether the detector functions as a gate or a trigger.

Quantized sensory input

To apply the aforementioned models on the random dot motion detection and discrimination task, we need to compute the log posterior ratios between pairs of classes (**Eq. 5.1** and **Eq. 5.2**). We chose a probabilistic strategy based on a front-end of direction tuning neurons. The front-end converts a visual stimulus (a length-Δ video segment of dots moving in space) into a set of action potentials, which are interpreted probabilistically to produce the log posterior ratios, as we see below.

$$
\lambda_\theta^k \triangleq \lambda_{min} + (\lambda_{max} - \lambda_{min}) \exp\left(-\frac{1}{2}\frac{\|\theta_k - \theta\|^2}{\sigma_Y^2}\right), \tag{5.6}
$$

where λ_{max} and λ_{min} are the maximum and minimum firing rates of a neuron (in Hz), and σ_Y is the tuning width, and the notation $\|\theta_k - \theta\|$ indicates the minimal angular distance between θ_k and θ.

> Here we have chosen a Gaussian tuning curve. This choice is not critical, e.g. a von Mises function [25] works equally well.

Within a unit interval $((t-1)\Delta, t\Delta]$, the number of spikes $X_{i,t}^k$ emitted from neuron k at location i in response to motion θ is (**Fig. 5.3b**):

$$
P(X_{i,t}^k = n) = Poisson(n|\lambda_\theta^k \Delta). \tag{5.7}
$$

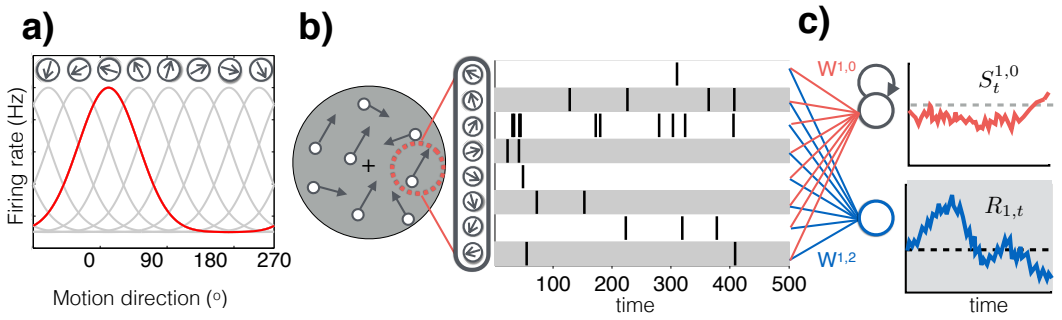

Figure 5.3: **From front-end MT neurons to log posterior ratios**. (**a**) Turing curves of a hypercolumn of 8 MT motion-tuning neurons (**Eq. 5.6**). (**b**) A hypercolumn of neurons activate in response to random dot motion at their receptive field (red dashed circle). Raster plot shows simulated homogeneous Poisson spike trains (**Eq. 5.7**) with max rate $\lambda_{max} = 20Hz$. (**c**) Two downstream neurons compute log posterior ratios from the spikes. $S_t^{1,0}$, the log posterior ratio between motion 1 and incoherent motion, may be computed by adding $W^{1,0}$-weighted spikes to a recurrent unit (**Eq. 5.10**). $R_{1,t}$, the log posterior ratio motion 1 and 2 after change onset, is a linear combination of the spike trains weighted by $W^{1,2}$ (**Eq. 5.13**). $W^{1,0}$ and $W^{1,2}$ depend on the motion coherence and motion directions, and are given in closed-form (**Eq. 5.9** and **Eq. 5.12**).

Log posterior ratios for detecting coherent motion from spikes

Consider a visual display with M moving dots. The dots are spaced sufficiently far apart such that each dot is monitored by a unique hypercolumn. At any point in time, a random fraction z ('coherence', $0 \leq z \leq 1$) of the M dots are moving along the same direction, and the remaining along random directions. Let $\bar{\lambda} \triangleq \mathbb{E}_\theta[\lambda_\theta^k]$ be a neuron's average firing rate over all stimulus directions. $\bar{\lambda}$ should be roughly identical for all neurons thanks to symmetry.

The log likelihood ratio $r_t^{c,0}$ between z fraction of coherent motion along direction θ_c of class $c \in \{1, 2\}$ and incoherent motion (class 0) is given by (derived in Methods **Eq. A.37**):

$$r_t^{c,0} \triangleq \log \frac{P(X_t|C_t = c)}{P(X_t|C_t = 0)} = \sum_i \sum_k W_k^{c,0} X_{i,t}^k, \tag{5.8}$$

$$\text{where} \quad W_k^{c,0} \triangleq \log \frac{(1-z)\bar{\lambda} + z\lambda_{\theta_c}^k}{\bar{\lambda}}. \tag{5.9}$$

The log posterior ratio $S_t^{c,0}$ between coherent motion in θ_c and incoherent motion may be computed recursively as (see Methods **Eq. A.46** for detailed derivations)

$$S_t^{c,0} = \mathcal{S}\text{rec}\left(S_{t-1} - \log \alpha_t\right) + \log \frac{\alpha_t}{1-\alpha_t} + r_t^{c,0}, \tag{5.10}$$

where $\mathcal{S}\mathrm{rec}\,(x) \triangleq \log(1+\exp(x)) \approx \max(0,x)$ is the 'soft-rectifier' function and α_t is the probability of a change happening now knowing that it has definitely not happened prior to t: $\alpha_t \triangleq P(t_\delta = t | t_\delta \geq t)$. The initial condition is $S_0 \triangleq \log \frac{P(t_\delta \leq 0)}{P(t_\delta > 0)} = \log 0 = -\infty$. See figure **Fig. 5.4** for an example of $S_t^{c,0}$.

The hardmax approximation gives an intuition for $S_t^{c,0}$. If past evidence util $t-1$ suggests that the likelihood for coherent motion is so low that $S_{t-1}^{c,0} - \log \alpha_t < 0$, then the system should "forget" about past evidence and reset to the log prior ratio $\log \frac{\alpha_t}{1-\alpha_t}$ instead. $\log \alpha_t$ thus is a threshold for triggering the forgetting mechanism. The forgetting mechanism allows $S_t^{c,0}$ to discard noisy observations in the distant past while taking in new evidence into consideration. For example, in **Fig. 5.4** we simulate the log posterior ratio $S_t^{c,0}$ for exponentially distributed change time. $S_t^{c,0}$ behaves almost like a memoryless system before the change occurs, which is crucial for an organism to detect changes whose arrival time spans a long duration.

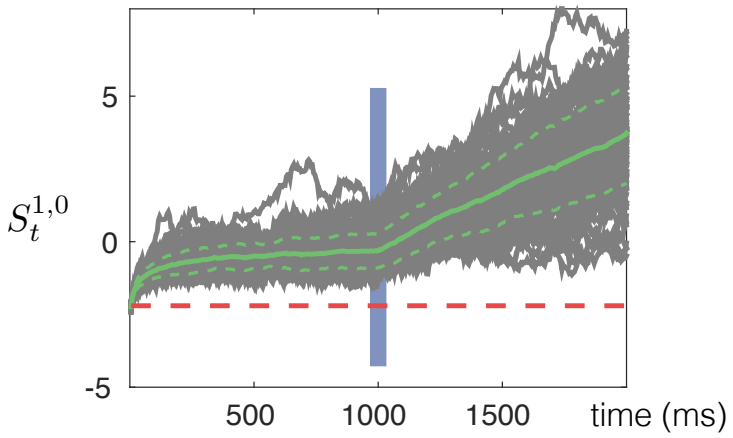

Figure 5.4: **The log posterior ratio $S_t^{c,0}$ for detecting coherent motion**. Log posterior ratio $S_t^{1,0}$ over 100 trials of Monte-Carlo simulations with mean and standard deviation overlaid. Unbeknown to the model, stimulus onset in all trials is $t_\delta = 1$ sec. The model instead uses an exponential prior for t_δ. The dash line shows the log prior ratio $\log \frac{\alpha_t}{1-\alpha_t}$. As we see from **Eq. 5.10**, the log prior ratio gives a lower bound for $S_t^{1,0}$ and cues the observer when to 'pay attention' and when to let go the past. $r_t^{1,0}$ is modeled by a Gaussian random walk $r_t^{1,0} \sim \mathcal{N}(\mu_C \Delta, \sigma^2 \Delta)$ with $\mu_C = \pm 14$ and $\sigma = 3.5$.

Log posterior ratios for classifying coherent motion from spikes

Using the same analysis we can compute the log likelihood ratio $r_t^{1,2}$ of the observation at time t between two different directions of coherent motion, θ_1 and θ_2, at

coherence level z.

$$r_t^{1,2} \triangleq \log \frac{P(X_t|C_t=1)}{P(X_t|C_t=2)} = \sum_i \sum_k W_k^{1,2} X_{i,t}^k, \tag{5.11}$$

$$\text{where} \quad W_k^{1,2} \triangleq \log \frac{(1-z)\bar{\lambda} + z\lambda_{\theta_1}^k}{(1-z)\bar{\lambda} + z\lambda_{\theta_2}^k}. \tag{5.12}$$

The log posterior ratio between coherent motions conditioning on post-change evidence is (derived in Methods **Eq. A.48**):

$$R_{t,t'} = \sum_{i=t}^{t'} r_i^{1,2}. \tag{5.13}$$

We assume even class priors $P(C=1) = P(C=2)$. Uneven prior may be incorporated by a simple shift of $R_{t,t'}$. See figure **Fig. 5.3c** for an example of $R_{t,t'}$.

> The expressions of the log posterior ratios in **Eq. 5.10** and **Eq. 5.13** suggest straightforward spiking implementations. The mechanisms are similar to those discussed in **Sec. 3.6**.

5.3 Psychophysics

Design

To test which of the proposed models (CD, DCP, DCS) best matches human detection and discrimination behavior, we recruited human subjects to participate in two experiments. Both experiments employed a dynamic random-dot display [2], where white dots were randomly distributed on a black background. All dots moved randomly except for a random fraction z that moved along a consistent direction. The direction could be one of two directions θ_1 and θ_2. The average of the two directions was always 90°, so we chose to represent them using the direction discrepancy $\Delta\theta = |\theta_2 - \theta_1|$. See **Fig. 5.1a**-c.

> Details of the display: the random dots with a density of $16.7 dots/deg^2/s$ were displayed with a 5° diameter circular aperture about the fixation center. Each dot was a white square of 5×5 pixels (0.14°). For the stimulus, on each video frame the coherently moving dots were shifted 0.125° from their positions $25ms$ earlier (three video frames, refresh rate = $120Hz$), corresponding to a speed of $5°/s$, while others were randomly repositioned.

The first experiment was dual detection and classification (**Fig. 5.1a**). The stimulus motion started incoherent ($z = 0$) and changed to one of two coherent directions

($z > 0$) after a stochastic delay t_δ. t_δ followed an exponential distribution with a mean of $800ms$. Subjects indicated the direction of coherent motion by button-press. Responses earlier than t_δ were considered false detections. Subjects were instructed to minimize both misclassification errors and response time while maintaining the false detection rate below 20%.

The second experiment was pure detection. With the identical setup as the dual experiment, here subjects were instructed to press a button as soon as they perceived coherent motion regardless of motion direction. The goal was to minimize response time while keeping the false detection rate below 20%.

We systematically varied the coherence level z and the direction discrepancy $\Delta\theta$ for each experiment. z is chosen randomly from $\{1.6\%, 3.2\%, 6.4\%, 12.8\%, 25.6\%\}$ and $\Delta\theta$ from $\{180°, 60°\}$. Both z and $\Delta\theta$ were fixed within a block of consecutive trials and varied between blocks (i.e. the subjects know the coherence level z and $\Delta\theta$).

Model fitting

We fit each model to the data that was collected in both experiments. The models were parameterized by (1) signal-to-noise ratio of the front-end and (2) decision thresholds. The front-end had four parameters: the minimum and maximum firing rates λ_{min} and λ_{max}, the tuning width w, and the number of neurons N. We fixed $\lambda_{min} = 1Hz$ and $w = 25°$ according to their physiological values in the macaque monkey [22]. Since N and λ_{max} have similar effects on the signal-to-noise ratio we did not fit both; rather, we fixed $N = 16$ neurons, and only fit the maximum firing rate λ_{max}. (2) CD only has one decision threshold τ_{dis}, whereas both DCP and DCS have two thresholds, τ_{det} and τ_{dis}, for their detector and classifier components, respectively.

We selected λ_{max} and the threshold(s) of each model and each subject to maximize the model prediction's agreement with the data in terms of the median response time, the misclassification rate and the false detection rate. The same λ_{max} parameter was used across different "conditions", parameterized by the coherence level z, the motion discrepancy $\Delta\theta$ and the experiment type (dual versus detection only). This parameter-sharing was made possible by the generalizability of our front-end model. By contrast, we did not share the thresholds, yielding one threshold in CD and two in DCP and DCS for each condition.

Human decision-making involves a perceptual component (evidence accumulation

to decision threshold) and a non-perceptional delay (axonal propagation, motor delays, etc). Our model only accounts for the perceptual component. The non-perceptional component was modeled phenomenologically with a log-normal distribution with two additional parameters (mean and variance) per subject.

Fitting results

The fits for the response time, misclassification and detection errors of a randomly sampled subject are shown in **Fig. 5.5** and **Fig. 5.6**. All three models qualitatively explain subjects' performance, although DCS is the most faithful to the error data. Despite the parsimonuous parameterization, the models predict the key performance metrics (**Fig. 5.5**) as well as the full response time histograms (**Fig. 5.6**) of the subjects. The overall scores of fitting the dual task and for fitting both tasks are in shown **Fig. 5.8a** and b. Both plots show significantly higher fitting errors of the optimal CD model compared to the sub-optimal models, within which the DCS perform better. This trend is also consistent across all 10 subjects except one (subject JD).

To further separate the two sub-optimal models, we also visualized the posterior estimates of the parameters in **Fig. 5.7.** The posterior weights the parameter values according to their agreement with the data. The signal-to-noise parameter λ_{max} estimate is correlated with the mean non-perceptual delay. This is not surprising as higher signal-to-noise ratio means shorter perceptual times, which leaves a shorter time to be explained by the non-perceptual delay. We see qualitatively that DCP produces less consistent estimates between the two experiments than does DCS. A quantitative comparison in **Fig. 5.8c** confirms that DCS is significantly more consistent across all subjects.

5.4 Discussion and summary

Plausible model for human behavior

We proposed three candidate mechanisms for joint detection and discrimination, and we explored whether any of them can account for human performance. Our first observation is that the optimal model CD underperforms the sub-optimal models in explaining human behaviors. This trend is significant and consistently observed in the dual tasks and joint fitting of both the dual and the detection tasks. However, the discrepancy between CD and the other models may be a consequence of the degree of freedom (DoF), as CD only has one threshold while DCP and DCS each independently manipulate two thresholds. To remove DoF as a confound, we re-

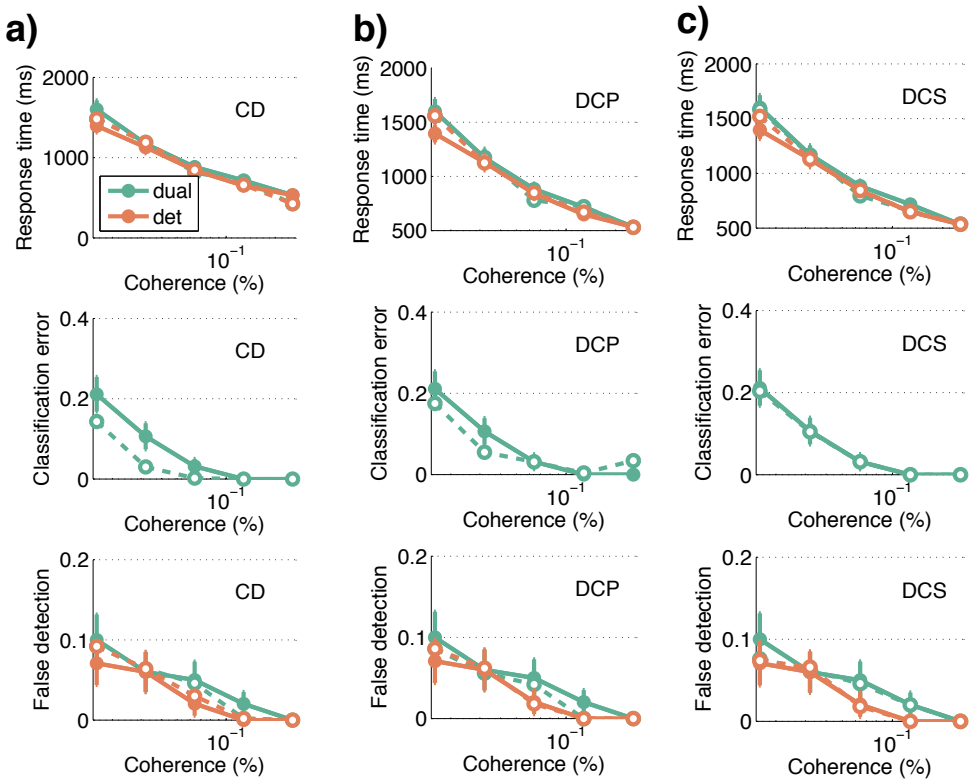

Figure 5.5: **Fitting results for a randomly selected subject**. The median response times, misclassification errors and false detection rates of a random subject (BW) and the fitted model predictions in the dual task and the detection task. The columns represent the three methods (CD, DCP and DCS). Solid lines show subject's data with 1 ste and dashed lines show the predictions. The direction discrepancy is $\Delta\theta = 180°$.

fitted the experiments while manipulating the parameter sharing across conditions. Even with more free parameters per condition, CD is less consistent with the data than the sub-optimal models are. Therefore, CD may not be the strategy of choice for humans.

DCS and DCP have the same DoF, hence may be compared fairly. The data suggests that DCS may be closer to the strategy for humans. A first clue is that DCS outperforms DCP in both fitting experiments across all subjects (**Fig. 5.8a** and b). A second cue comes from comparing the posteriors (**Fig. 5.7**). In the pure detection experiment, DCS and DCP reduce to the same algorithm, hence their parameter estimates should also be *the same*. In the dual task, however, DCP and DCS should produce *different* estimates, as explained below. DCS discards information prior to detector activation and thus requires longer evidence accumulation to achieve the

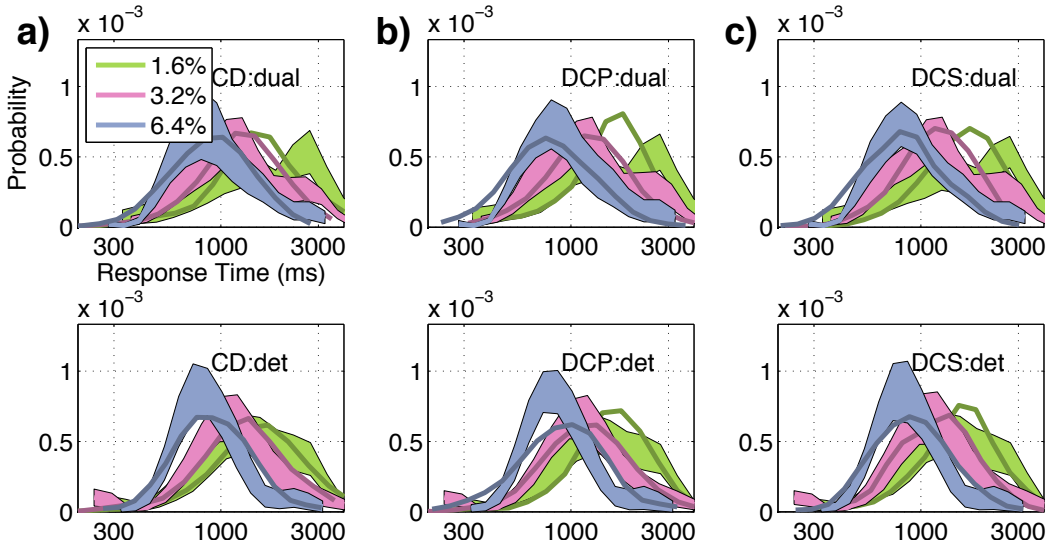

Figure 5.6: **Fitting results for a random subject (cont'd)**. The response time histograms and fits for the dual task (first row) and the detection task (second row) of a randomly selected subject (BW). Only a subset of the data, with coherence levels $\{1.6\%, 3.2\%, 6.4\%\}$ and direction discrepancy $\Delta\theta = 180°$, are shown. Each color denotes a different coherence level. The solid lines are fits from the model and the filled regions show the mean ± 1 bootstrapped standard error of the subject's data. Each column shows the fit of one model (CD, DCP and DCS).

same level of accuracy as does DCP. Therefore, to explain the same data DCS must compensate with a higher λ_{max} estimate, a shorter motor delay estimate, or both. As a result, comparing the posteriors between the two tasks will expose the incorrect model. In the case of **Fig. 5.7a**nd **Fig. 5.8c**, we see that DCS produces consistent parameter estimates, while DCP does not and should therefore be eliminated.

Sub-optimal information processing

Our analysis suggests that humans are sub-optimal in the dual detection-decision task; however, this conclusion is not inconsistent with previous findings [11], [17]–[19] that the human visual system is near-optimal in evidence accumulation. In the DCS model sub-optimality resides in the decision strategy rather than in evidence accumulation.

We speculate that the human visual system may use the DCS strategy for two reasons. 1) Modularity. DCS may be adapted to tackle a pure detection tasks by simply setting the classification threshold τ_{dis} to zero. Similarly, setting $\tau_{det} = -\infty$ would tune DCS for a pure classification task. The flexibility to switch between tasks is a desirable property. For instance, sound cues may sometimes permit detection,

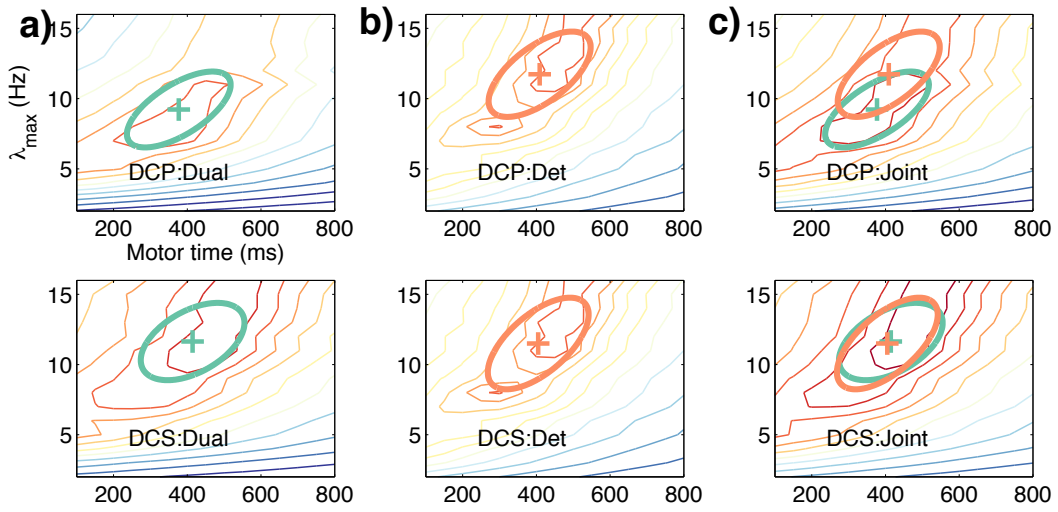

Figure 5.7: **Posterior distribution of parameters for a random subject**. The posterior of the signal-to-noise parameter (λ_{max}) and the non-perceptual delay parameter (motor time) for DCP (first row) and DCS (second row) for a random subject (BW). The three columns represent posteriors obtained from (**a**) the dual task, (**b**) the detection task, and (**c**) both tasks. For each panel in (a) and (b) the ellipse and the cross represent a Gaussian approximation to the posterior and its mean. In (c) the two ellipses from (a) and (b) are superimposed.

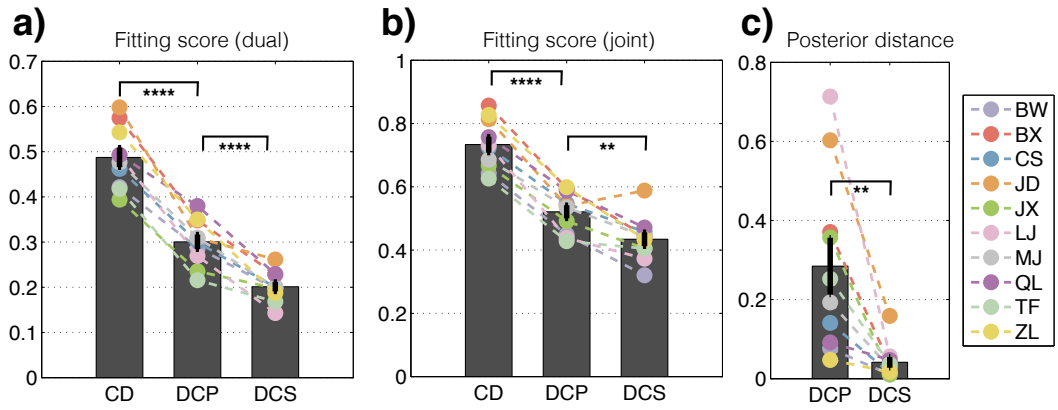

Figure 5.8: **Fitting performance**. (**a**) Fitting scores (lower means better) of the three models CD, DCP and DCS on the dual task. (**b**) Same as (a) except that the score is computed jointly over the dual and the detection-only task. (**c**) Distances (lower means better) between the posterior obtained from the dual task and that from both tasks. Colored dashed lines show performance for different subjects (see legend). Bars show average performance over 10 subjects with 1 ste. '**' represents $p \leq 0.01$ and '****' represents $p < 10^{-4}$.

which renders visual detection unnecessary. As a result, the visual system would need to switch from the dual task mode to pure classification mode. 2) Power

efficiency. While the classifiers in CD are bombarded with sensory inputs all the time, the classifier in DCS only activates for brief moments when a change in the environment has been confirmed. In other words, the classifier in DCS may be dormant most of time to conserve energy. This advantage may be more pronounced for discrimination tasks with a large number of categories, as the relative energy reduction from CD to DCS is proportional to the number of categories. On the other hand, in situations when changes happen frequently and only a small number of classes are involved, CD may be a viable strategy for humans.

References

[1] W. S. Geisler, "Sequential ideal-observer analysis of visual discriminations.," *Psychological Review*, vol. 96, no. 2, pp. 267–314, 1989.

[2] W. T. Newsome, K. H. Britten, and J. A. Movshon, "Neuronal correlates of a perceptual decision.," *Nature*, 1989.

[3] J. Palmer, A. C. Huk, and M. N. Shadlen, "The effect of stimulus strength on the speed and accuracy of a perceptual decision," *Journal of Vision*, vol. 5, no. 5, pp. 376–404, 2005.

[4] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[5] R. Ratcliff, "Theoretical interpretations of the speed and accuracy of positive and negative responses.," *Psychological Review*, vol. 92, no. 2, p. 212, 1985.

[6] D. J. Lasley and T. Cohn, "Detection of a luminance increment: Effect of temporal uncertainty," *Journal of the Optical Society of America*, vol. 71, no. 7, pp. 845–850, 1981.

[7] H. Pashler, "Familiarity and visual change detection," *Perception & Psychophysics*, vol. 44, no. 4, pp. 369–378, 1988.

[8] J. Hohnsbein and S. Mateeff, "The time it takes to detect changes in speed and direction of visual motion," *Vision Research*, vol. 38, no. 17, pp. 2569–2573, 1998.

[9] C. W. Clifford and M. Ibbotson, "Fundamental mechanisms of visual motion detection: Models, cells and functions," *Progress in Neurobiology*, vol. 68, no. 6, pp. 409–437, 2002.

[10] E. P. Cook and J. H. Maunsell, "Dynamics of neuronal responses in macaque mt and vip during motion detection," *Nature Neuroscience*, vol. 5, no. 10, pp. 985–994, 2002.

[11] C. M. Glaze, J. W. Kable, and J. I. Gold, "Normative evidence accumulation in unpredictable environments," *Elife*, vol. 4, e08825, 2015.

[12] W. Shewhart, "The application of statistics as an aid in maintaining quality of a manufactured product," *Journal of the American Statistical Association*, vol. 20, no. 152, pp. 546–548, 1925.

[13] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[14] Y. Ritov, "Decision theoretic optimality of the cusum procedure," *The Annals of Statistics*, pp. 1464–1469, 1990.

[15] M. Pollak and D. Siegmund, "Approximations to the expected sample size of certain sequential tests," *The Annals of Statistics*, pp. 1267–1282, 1975.

[16] G. Lorden, "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, pp. 1897–1908, 1971.

[17] W. J. Ma, V. Navalpakkam, J. M. Beck, R. Van Den Berg, and A. Pouget, "Behavior and neural basis of near-optimal visual search," *Nature Neuroscience*, vol. 14, no. 6, pp. 783–790, 2011.

[18] B. Chen and P. Perona, "Speed versus accuracy in visual search: Optimal performance and neural architecture," *Journal of Vision*, vol. 15, no. 16, pp. 9–9, 2015.

[19] J. Drugowitsch, G. C. DeAngelis, D. E. Angelaki, and A. Pouget, "Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making," *ELife*, vol. 4, e06678, 2015.

[20] E. P. Simoncelli, L. Paninski, J. Pillow, and O. Schwartz, "Characterization of neural responses with stochastic stimuli," *The Cognitive Neurosciences*, vol. 3, pp. 327–338, 2004.

[21] M. Jazayeri and J. A. Movshon, "Optimal representation of sensory information by neural populations," *Nature Neuroscience*, vol. 9, no. 5, pp. 690–696, 2006.

[22] J. H. Maunsell and D. C. Van Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation," *Journal of neurophysiology*, vol. 49, no. 5, pp. 1127–1147, 1983.

[23] M. Shadlen and W. Newsome, "Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey," *Journal of Neurophysiology*, vol. 86, no. 4, pp. 1916–1936, 2001.

[24] J. Drugowitsch, R. Moreno-Bote, A. Churchland, M. Shadlen, and A. Pouget, "The cost of accumulating evidence in perceptual decision making," *The Journal of Neuroscience*, vol. 32, no. 11, pp. 3612–3628, 2012.

[25] B. Amirikian and A. P. Georgopulos, "Directional tuning profiles of motor cortical cells," *Neuroscience Research*, vol. 36, no. 1, pp. 73–79, 2000.

Chapter 6

OPTIMALITY ANALYSIS OF SEQUENTIAL PROBABILITY RATIO TEST

Strictly Optimal Sequential Tests

The sequential probability ratio test (SPRT) is *asymptotically optimal* in the speed versus accuracy tradeoff (SAT) for problems such as visual search (**Ch. 3**) and scotopic object recognition (**Ch. 4**), but how close to optimal is SPRT in the *non-asymptotic case*, i.e. when the cost of error η or the expected response time is small? We numerically compare SPRT and the optimal strategy on the homogeneous visual search (**Sec. 3.4**) problem and propose alternative test forms that may be optimal in non-asymptotic scenarios.

6.1 Optimal decision strategy for homogeneous search

Recall that the goal of homogeneous visual search is to detect whether a target appears anywhere in a field of display ($C = 1$ if target present, and $C = 0$ otherwise). All locations contain either a target or a distractor, and at most one target appears at a time. The target may be separated from a distractor using unique features (orientation). The observations are the action potentials $\boldsymbol{X}_{1:t} = \{\boldsymbol{X}^l_{1:t}\}_{l=1}^M$ V1 orientation-tuned hypercolumns from all M display locations.

A decision strategy for homogeneous visual search aims to minimize Bayes risk (**Eq. 2.1**):

$$\text{Risk} = \mathbb{E}[T] + \eta\mathbb{E}[\hat{C}_T \neq C],$$

where $\hat{C}_T \in \{0, 1\}$ is the observer's decision at decision time T, η is the relative cost of error with respect to time. The optimal test achieves the lowest risk among all tests.

> For simplicity we assume that false positives and false negatives have the same cost, and so do the response times under each class. Different costs can be easily accommodated without affecting the overall analysis.

Two components are necessary to describe the optimal test: a state space $\boldsymbol{Z}(t)$ over time and a decision strategy that associates each state and time with an action. One

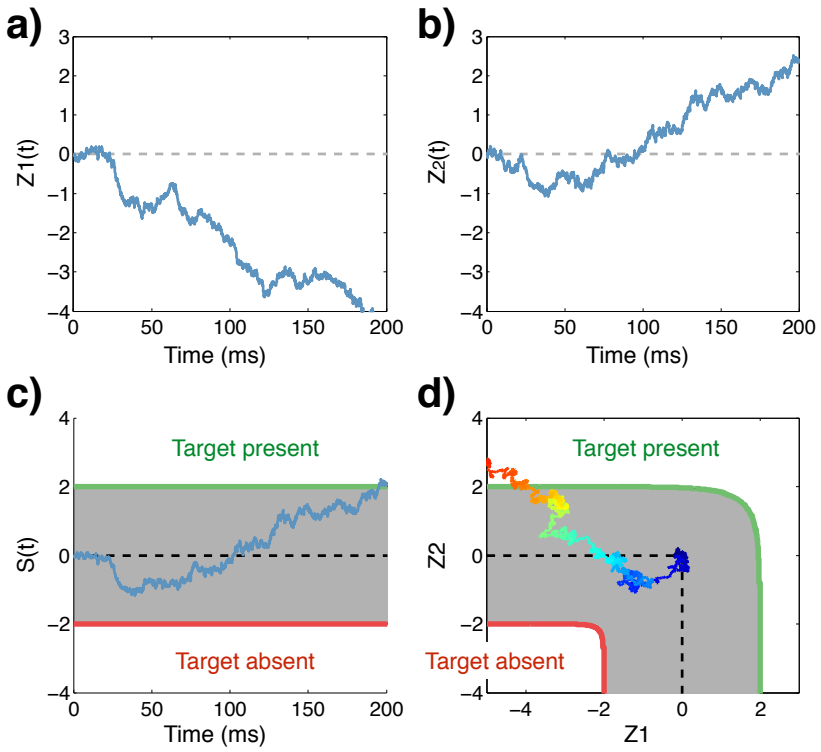

Figure 6.1: **Decision strategies for homogeneous visual search.** To perform probabilistic inference, a sequential test computes for each location the local log likelihood ratio $Z_l = \log \frac{P(X^l|C_l=1)}{P(X^l|C_l=0)}$ over time: **(a)** Z_l at a distractor location, **(b)** Z_l at the target location. **(c,d)** Two decision strategies that make use of the probabilistic interpretation for a two-dimensional visual search problem. SPRT **(c)** thresholds the one-dimensional log likelihood ratio $S(\boldsymbol{X}_{1:t})$ (**Eq. 3.7**), whereas the optimal **(d)** uses a decision boundary in the joint space of $\{Z_1, Z_2\}$. Time in (d) is color-coded, cooler colors means earlier.

common constraint on the state space is that it must be Markov in time: $\boldsymbol{Z}(t)$ must be sufficient in summarizing past observations so that given $\boldsymbol{Z}(t)$, future observations become independent from the past (see Appendix **Sec. A.4**). Once this constraint is satisfied, the problem may be formulated as a partial observation Markov decision process (POMDP)[1], and the optimal strategy may be solved exactly using dynamic programming.

We choose $\boldsymbol{Z}(t)$ to be the collection of log posterior ratios from all locations: $\boldsymbol{Z}(t) = \{Z_l(t)\}_{l=1}^{M}$ and

$$Z_l(t) \triangleq S(\boldsymbol{X}^l_{1:t}). \tag{6.1}$$

For simplicity we consider the most common formulation of input as a Gaussian random walk at each location (e.g. [2], [3]). This approximates the Poisson model

used in **Ch. 3**, which is more expensive to simulate. The input is parameterized by the *drift-rate* $\mu_{C,l}$, which depends on the stimulus class C and the location l (**Fig. 6.1a,c**). A larger drift-rate difference between the two classes $|\mu_{1,l} - \mu_{0,l}|$ implies a higher signal-to-noise ratio, or equivalently, an easier discrimination problem at location l.

Computational solution for low-dimensional problems

The optimal decision may be computed numerically using dynamic programming [1], [4]. Define $R(\boldsymbol{Z}(t), t)$ as the lowest total risk an observer could incur starting from state $\boldsymbol{Z}(t)$ at time t. The optimal risk is equivalent to $R(\vec{0}, 0)$, the total risk from time 0 onwards with a flat prior. $R(\boldsymbol{Z}, t)$ is recursively given by:

$$R(\boldsymbol{Z}(t), t) = \min \begin{cases} \eta(1 - P_0(\boldsymbol{Z}(t))) & D = 0\text{: declare target absent} \\ \eta P_0(\boldsymbol{Z}(t)) & D = 1\text{: declare target present} \\ \Delta + \mathbb{E}_{\boldsymbol{Z}(t+\Delta)|\boldsymbol{Z}(t)} R(\boldsymbol{Z}(t+\Delta), t+\Delta) & D = \emptyset\text{: wait.} \end{cases} \tag{6.2}$$

At any time t and any state $\boldsymbol{Z}(t)$, the ideal observer picks the action $D \in \{\emptyset, 0, 1\}$ that yields the lowest risk. If declaring target-absent, the observer makes a false rejection mistake. The false reject probability can be computed from the state $\boldsymbol{Z}(t)$ and is denoted $P_0(\boldsymbol{Z}(t))$ (see Appendix **Sec. A.4**). If waiting for more evidence, the observer trades off the cost $C_{\text{time}}\Delta$ for a new observation of duration Δ, and access to the cumulative risk at time $(t + 1)$.

The optimal decision strategy is defined over an $M + 1$ dimensional state-space. The state space is separated by decision boundaries/surfaces into three different decision regions [5]. Furthermore, the recurrence equation 6.2 is time invariant. As a result, the optimal decision is constant in time (see [1]) and the decision surfaces have $M - 1$ dimensions.

Conjecture for high-dimensional problems

Recall that the optimal decision strategy for homogeneous visual *discrimination* (between two simple alternatives), is SPRT. We conjecture that the optimal decision strategy for homogenous visual *search* is similar to SPRT: it uses two SPRTs defined on scaled log posterior ratios.

Conjecture 1 *(Uniform drift-rates) If all locations share the same drift-rate ($\mu_{1,l} = -\mu_{0,l} = \mu, \forall l$), let τ_+ and τ_- be the optimal upper and lower thresholds for visual discrimination at location l associated to a cost of error of η, then the optimal*

decision surfaces for homogeneous visual search with the same cost of error η are:

$$S_+(\mathbf{Z}(t)) = \frac{1}{a_+} \underset{l=1,\ldots,M}{\mathcal{S}max} \left(a_+(Z_l(t) - \log(M))\right) \geq \tau_+, \tag{6.3}$$

$$S_-(\mathbf{Z}(t)) = \frac{1}{a_-} \underset{l=1,\ldots,M}{\mathcal{S}max} \left(a_-(Z_l(t) - \log(M))\right) \leq \tau_-, \tag{6.4}$$

where a_+ and a_- are unknown parameters.

Conj. 1 states that the optimal decision strategy is to wait until either $S_+(X(t)) \geq \tau_+$ to declare $\hat{C} = 1$ or $S_-(X(t)) \leq \tau_-$ to declare $\hat{C} = 0$. The thresholds τ_+ and τ_- are obtained easily by solving a one-dimensional dynamic programming problem [3]. The thresholds are chosen to guarantee asymptotic optimality. Intuitively, when there is only one location ($M = 1$), the problem reduces to visual discrimination and **Conj. 1** reduces to SPRT, which is optimal for visual discrimination. For $M > 1$, asymptotically one "winner" will emerge from the M locations, and $Z_l(t)$ at other locations become negligible compared to that of the winner location l^*. The decision is effectively reduced to concerning only the winner location l^*. In this case:

$$S_+(\mathbf{Z}(t)) = \frac{1}{a_+} \underset{l=1,\ldots,M}{\mathcal{S}max} \left(a_+(Z_l(t) - \log(M))\right) \approx Z_{l^*}(t) - \log(M),$$

$$S_-(\mathbf{Z}(t)) \approx Z_{l^*}(t) - \log(M).$$

Any location could be the winner location with a probability $1/M$, hence asymptotically the visual search problem reduces to a visual discrimination problem at location l^* with a log prior ratio of $\log(1/M)$. This reduced problem may be solved optimally using adjusted thresholds $\tau_+ + \log(M)$ and $\tau_- + \log(M)$ (for proof see Appendix **Sec. A.4**), which matches the asymptotic behavior of the conjecture.

Fig. 6.2(a-b) and **Fig. 6.3** show excellent empirical match between the conjectured thresholds and the optimal thresholds in 2D.

Our conjecture can be extended to cases where the drift-rates are different across locations.

Conjecture 2 *(Non-uniform drift-rates) Let $\tau_+^{(l)}$ and $\tau_-^{(l)}$ be the optimal upper and lower thresholds for visual discrimination at location l associated with a cost of error of η, define $c_+^{(l)} = \tau_+^{(M)}/\tau_+^{(l)}$ and $c_-^{(l)} = \tau_-^{(M)}/\tau_-^{(l)}$, the optimal decision surface*

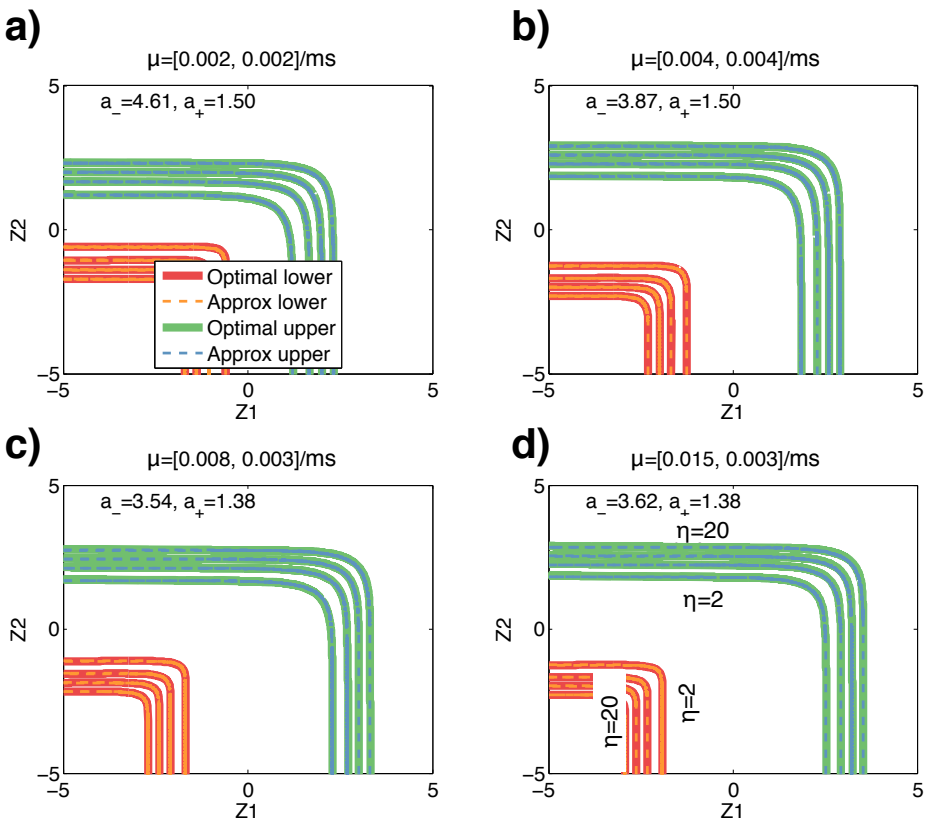

Figure 6.2: **Optimal sequential test for 2D visual search.** **(a-b)** Optimal decision thresholds and approximations for different costs of errors $\eta =\in \{2, 5, 10, 20\}$) in homogeneous search. Decision boundaries are approximated using **Eq. 6.5** and 6.6 with $a_+ = 1.50$ and $a_- = 4.61$. **(c-d)** Optimal decision thresholds and approximations for heterogeneous drift-rate search. Drift-rates are (a-b) $\pm 2/sec$, (c) $\{\pm 8, \pm 3\}/ms$ and (d) $\{\pm 15, \pm 3\}/sec$.

for visual search with the same time cost is:

$$S_+(\mathbf{Z}(t)) = \frac{1}{a_+} \underset{l=1,\ldots,M}{\mathcal{S}max} \left(c_+^{(l)} a_+ (Z_l(t) - \log(M)) \right) \geq \tau_+^{(M)}, \tag{6.5}$$

$$S_-(\mathbf{Z}(t)) = \frac{1}{a_-} \underset{l=1,\ldots,M}{\mathcal{S}max} \left(c_-^{(l)} a_- (Z_l(t) - \log(M)) \right) \geq \tau_-^{(M)}. \tag{6.6}$$

Conj. 2 only differs from **Conj. 1** for uniform drift-rate (**Eq. 6.5**) in that the local diffusions are scaled by a location-dependent factor $c_+^{(l)}$. and $c_-^{(l)}$. These factors normalize the diffusion at each location by its efficiency. The normalization is with respect to a reference location, which is arbitrarily chosen to be location M. In the

asymptotic case where only one location l^* is relevant,

$$S_+(\boldsymbol{Z}(t)) \approx \tau_+^{(M)}(Z_{l^*}(t) - \log(M))/\tau_+^{(l^*)}, \tag{6.7}$$

$$S_-(\boldsymbol{Z}(t)) \approx \tau_-^{(M)}(Z_{l^*}(t) - \log(M))/\tau_-^{(l^*)}, \tag{6.8}$$

and the visual search problem reduces to visual discrimination at location l^*. Since $\tau_+^{(l^*)}$ and $\tau_-^{(l^*)}$ are the optimal thresholds for visual discrimination, visual search should be optimal when $S_+(\boldsymbol{Z}(t))$ reaches $\tau_+^{(l^*)}$ or when $S_-(\boldsymbol{X}(t))$ reaches $\tau^{(l^*)}$. Substituting **Eq. 6.8** we obtain thresholds $\tau_+^{(M)}$ for $S_+(\boldsymbol{Z}(t))$ and $\tau_-^{(M)}$ for $S_-(\boldsymbol{Z}(t))$ (**Eq. 6.5**).

Conj. 2 only requires solving M one-dimensional dynamic programming problems for $\tau_+^{(l)}$ and $\tau_-^{(l)}$, which is more scalable than the optimal procedure (**Eq. 6.2**) that scales exponentially with M. **Fig. 6.2**(c-d) shows that the predicted thresholds from **Conj. 2** match the optimal thresholds from dynamic programming in 2D for a variety of costs of time and drift-rates.

6.2 Optimality analysis of current search models

How are existing visual search strategies compare against the optimal? For fairness we compare only approaches that perform probabilistic inference on the graphical model in **Fig. 3.1b**. These approaches, listed below, differ only in the decision strategy [6]:

a-SPRT (**Fig. 6.1d**): our two-SPRT approach that uses two decision surfaces prescribed in **Conj. 1** and **Conj. 2** to approximate the ideal observer.

SPRT [7] (**Fig. 6.1b**): a Bayesian extension of Ward's SPRT [8] into testing composite hypotheses. SPRT compares the log likelihood ratio of target-present versus target-absent $S(\boldsymbol{X}_{1:t})$ (**Eq. 3.7**) against a pair of thresholds. Since the SPRT is subject to the same asymptotic analysis in **Conj. 1**, it uses the same thresholds τ_- and τ_+ as does the a-SPRT. Essentially, SPRT is a special case of **Eq. 6.3** and **Eq. 6.4** where $a_+ = a_- = 1$.

SPRT-opt: the same as SPRT above except that it optimizes the upper and lower thresholds to minimize the risk function (**Eq. 2.1**). Since SPRT-opt may use different thresholds from those in the regular SPRT, it may not be asymptotically optimal. However, this does not prevent SPRT-opt from outperforming the regular SPRT (which is asymptotically optimal). This is because the asymptotic (i.e. long) decisions may only take up a tiny fraction of all the decisions (especially in easy tasks), and SPRT-opt may do better by focusing on the risk for shorter decisions.

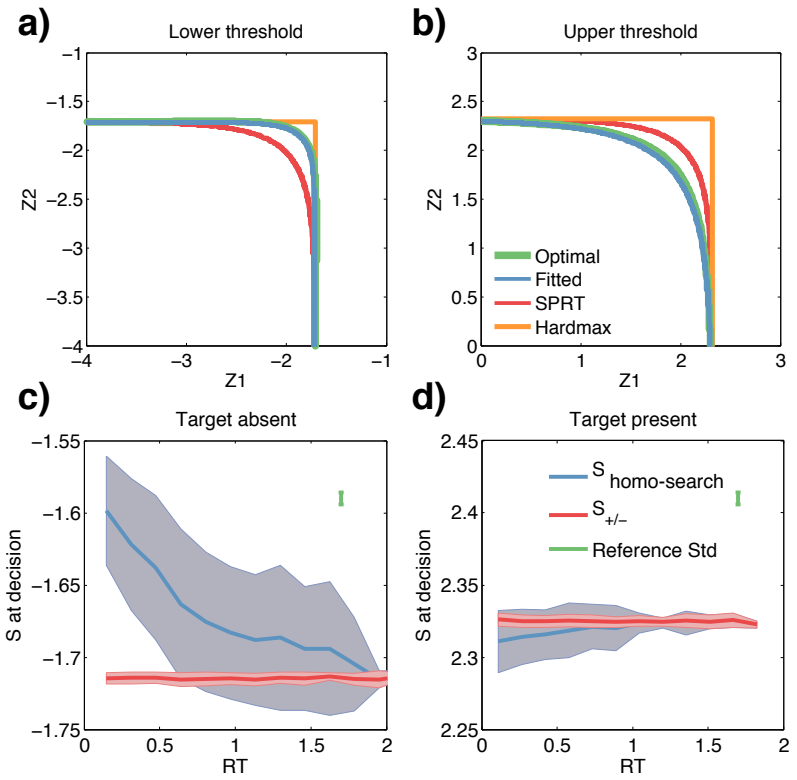

Figure 6.3: **Sequential testing strategies for homogeneous visual search in two-dimensions.** The optimal and various alternative decision strategies are compared in terms of **(a)** the lower and **(b)** the upper threshold in the joint space of $\{Z_1, Z_2\}$. The a-SPRT thresholds are obtained from **Eq. 6.3** and **Eq. 6.4** with $a_+ = 1.5$ and $a_- = 3.9$; both SPRT and Hardmax use the optimal threshold for visual discrimination so that asymptotically they are consistent with the optimal strategy. Input to each display location has a drift-rate of $\pm 4/sec$. **(c-d)** Each panel shows the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ distribution at the time of decision under the optimal decision strategy from $1k$ Monte-Carlo simulations. As references, the distribution of S_- when target is absent (c) and of S_+ when present (d) are shown. $S_\pm$ is not deterministic because time is discretized in the simulation, which causes the log likelihood ratios to have finite-sized jumps. Standard deviations of the jumps are shown as another reference. Drift-rate of the observation is $\pm 2/sec$.

Hardmax [7], [9]: an efficient approximation to SPRT. Each location decides whether it contains a target ($C^l = 1$) or a distractor ($C^l = 0$) based solely on the local belief $S(\boldsymbol{X}^l_{1:t})$. The observer declares target-present when any location reports a target detection, declares target-absent when all locations report a distractor, and waits for more information otherwise. Hardmax is also a special case of **Eq. 6.3** and **Eq. 6.4** where $a_+ = a_- = \infty$.

Decision surfaces comparison.

We want to see how these approaches differ from the optimal in various aspects. First, how different are their decision surfaces? In **Fig. 6.3**(a-b), we compare them on a visual search task with two display locations where it is computationally feasible to solve for the optimal decision boundary using dynamic programming. Since the decision boundaries are constant in time, they can be visualized in the 2-D space of Z_1 and Z_2 only. Each decision boundary is of the form $\{(Z_1, Z_2)|S(Z_1, Z_2) = \tau\}$, i.e. all pairs of Z_1 and Z_2 that could make the log likelihood ratio S reach a threshold of τ.

We observe that both the Hardmax and SPRT *differ significantly* from the optimal in terms of the decision surfaces (**Fig. 6.3**(a-b)). SPRT is conservative, because both thresholds bend outwards with respect to the optimal thresholds, which translates to longer decision times for both target-present and target-absent runs. Hardmax, on the other hand, is faster in declaring target-absent but slower in declaring target-present.

Can time-varying threshold make SPRT optimal?

A common practice in modeling decision making in visual discrimination is to employ a time-varying threshold. Can the optimal decision mechanism for visual search also be implemented using SPRT-opt with a *time-varying* threshold? We reject this hypothesis by computing the $S(\boldsymbol{X}_{1:t})$ distribution at the time of decision under the optimal test (**Fig. 6.3**(c-d)). If a time-varying threshold exists on $S(\boldsymbol{X}_{1:t})$ to recover the optimal strategy, the $S(\boldsymbol{X}_{1:t})$ values should be unique at the time of decision. Instead, we observe a wide spread in the $S(\boldsymbol{X}_{1:t})$ distribution. Therefore, $S(\boldsymbol{X}_{1:t})$ is not a sufficient statistic to implement the optimal test, and SPRT is sub-optimal in visual search [8].

Risk comparison.

The decision surfaces comparison above has one caveat: we consider all places on the decision boundary where decisions *could* be taken, ignoring the fact that some places on the boundary are more likely to be reached than others in an actual decision task. E.g., consider **Fig. 6.3b**, when the search task is easy, the diffusions when the target is present will most likely fall in the region of $\{Z_2 > 0, Z_1 \ll 0\}$ and $\{Z_2 \ll 0, Z_1 > 0\}$, and rarely visit the region of $\{Z_1 > 0, Z_2 > 0\}$ where the difference among the strategies is the most noticeable. This reasoning suggests that we should compare these strategies in terms of their actual *risk* value.

The risks for the strategies in a homogeneous search task are shown in **Fig. 6.4.**

Hardmax and SPRT are highly sub-optimal. SPRT-opt is almost indistinguishable from a-SPRT in the low time-cost scenario, but becomes sub-optimal when the cost of error becomes very small, i.e. when the decision time is short. Although we have not yet proven that a-SPRT is optimal, it is sufficient to conclude that any model that underperforms it is sub-optimal.

For search tasks where the drift-rates are non-uniform in space (**Fig. 6.5**), we see that even with two display locations, both SPRT-opt and Hardmax[1] are suboptimal when the drift-rates differ significantly across locations. The sub-optimality becomes progressively more pronounced as the heterogeneity of drift-rates increases. Behaviorally, when the drift-rate heterogeneity is large, Hardmax achieves near-identical ER vs RT trade-offs at both locations, whereas SPRT-opt and a-SPRT learn to sacrifice the ER at the low drift-rate location for a faster RT overall (**Fig. 6.5c**).

In conclusion, decision strategies employed by existing search models are sub-optimal. Hardmax, where one combines local decisions to reach a global decision, is sub-optimal in almost all scenarios. The SPRT-opt, where one executes a one-dimensional SPRT with optimized thresholds, is near-optimal in low cost, homogeneous search scenarios. When the cost of error is small and when the drift-rate is heterogeneous across locations, the SPRT-opt becomes sub-optimal, but remains similar to the optimal SATstrategy.

6.3 Chapter summary

We conjecture a novel procedure, a-SPRT, to compute the optimal decision strategy for high-dimensional visual search with uniform and non-uniform drift-rates in space. The a-SPRT makes use of two one-dimensional SPRTs with different scaling factors, and with thresholds that are constant in time. In two dimensions, the resultant decision boundary matches closely that of the optimal strategy. The conjecture is preferred over the standard dynamic programming procedure, which does not scale to high (more than three) dimensions.

We compare common models of visual search in their optimality in SAT. We discover that most of them are sub-optimal. While SPRT behaves similarly as the optimal strategy in homogeneous search tasks with uniform drift-rates, it is sub-optimal once the drift-rates become heterogeneous across locations.

[1] We do not include SPRT because it is not clear how to condense the M asymptotically optimal thresholds, one for each decision surface, into just one for the SPRT. Instead we trust that SPRT-opt, with the ability to optimize the thresholds, should always outperform any SPRT.

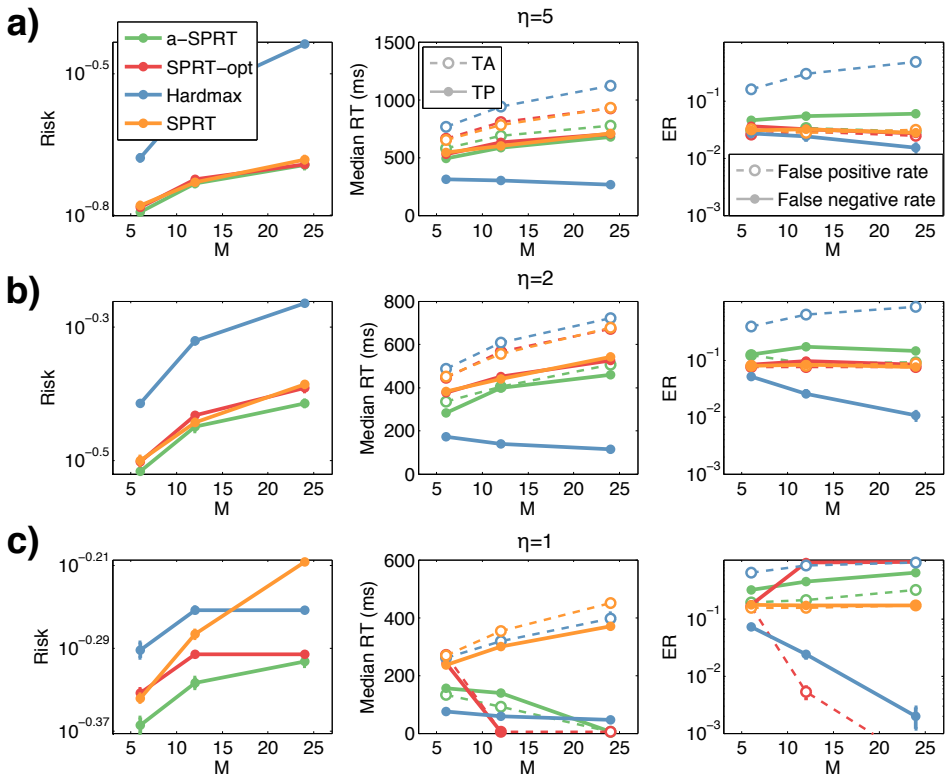

Figure 6.4: **Risk comparison of common decision strategies in homogeneous visual search.** a-SPRT, SPRT-opt, SPRT and Hardmax are compared under different costs of errors: **(a)** $\eta = 5$, **(b)** $\eta = 2$, and **(c)** $\eta = 1$ with a drift-rate of $\pm 12/sec$. Hardmax is sub-optimal in all cases. Regular SPRT is sub-optimal in the high cost scenario. SPRT-opt slightly under-performs a-SPRT in terms of the risk. a-SPRT and SPRT-opt are similar in terms of the RT during target-present (TP) and target-absent (TA), as well as the false positive rate and the false negative rate. Error bars are one standard error computed from $10k$ runs.

We highlight several unsolved issues for future work. First, it remains an open question why the optimal decision boundaries for homogeneous search can be described by two scaled-SPRTs. Second, we do not know how the scaling factors a_+ and a_- depend on search parameters, and therefore must search numerically for their values to minimize the risk. A better understanding is required to generalize ideal observers of visual search into greater dimensionality and heterogeneity. Third, in light of the marked difference between alternative models and the optimal strategy in the case of non-uniform drift-rates, it would be interesting to test subjects in this case to see which model best captures human behavior, and whether humans are optimal.

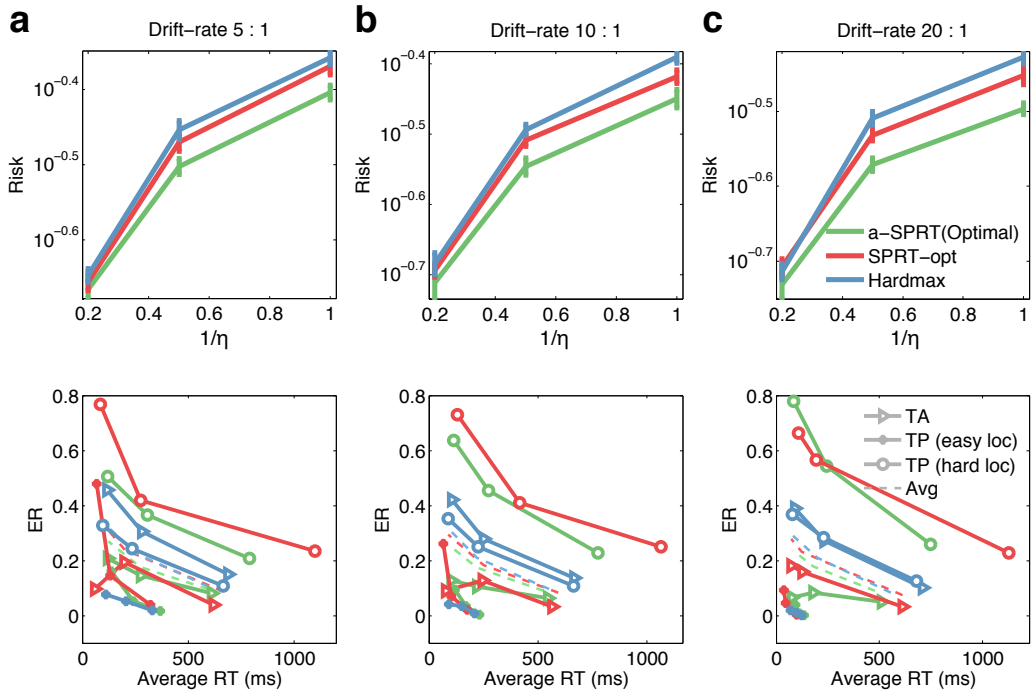

Figure 6.5: **Risk comparison of common decision strategies in heterogeneous drift-rate visual search.** a-SPRT, SPRT-opt and Hardmax are compared under various costs of time. The first row shows the overall risk versus the cost of error. The second row shows the ER vs RT tradeoff under different costs of errors (dots) and under three separate conditions (lines): target-absent (TA), target-present (TP) at the location with a larger drift-rate (easy) and target-present at the hard location. Drfit-rates are **(a)** $\{\pm 5, \pm 1\}/sec$, **(b)** $\{\pm 10, \pm 1\}/sec$ and **(c)** $\{\pm 20, \pm 1\}/sec$. One standard error in both RT and ER computed from $1k$ runs are shown but are too small to be visible. Both SPRT-opt and Hardmax underperform the optimal test.

References

[1] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman, "Acting optimally in partially observable stochastic domains," in *Association for the Advancements of Artificial Intelligence*, vol. 94, 1994, pp. 1023–1028.

[2] T. L. Thornton and D. L. Gilden, "Parallel and serial processes in visual search.," *Psychological Review*, vol. 114, no. 1, p. 71, 2007.

[3] J. Drugowitsch, R. Moreno-Bote, A. Churchland, M. Shadlen, and A. Pouget, "The cost of accumulating evidence in perceptual decision making," *The Journal of Neuroscience*, vol. 32, no. 11, pp. 3612–3628, 2012.

[4] R. Bellman, "Dynamic programming and lagrange multipliers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, no. 10, p. 767, 1956.

[5] M. Sobel *et al.*, "An essentially complete class of decision functions for certain standard sequential problems," *The Annals of Mathematical Statistics*, vol. 24, 1953.

[6] J. Palmer, P. Verghese, and M. Pavel, "The psychophysics of visual search," *Vision Research*, vol. 40, no. 10, pp. 1227–1268, 2000.

[7] B. Chen, V. Navalpakkam, and P. Perona, "Predicting response time and error rates in visual search," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

[8] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[9] J. Najemnik and W. S. Geisler, "Eye movement statistics in humans are consistent with an optimal search strategy," *Journal of Vision*, vol. 8, no. 3, p. 4, 2008.

Chapter 7

DISCUSSION AND CONCLUSIONS

The central thesis of this work is that the quantization of visual signals should be accounted for in vision algorithms. Information in the visual world does not become available all at once to an observer. Rather, it trickles in one quantum at a time: photons, action potentials, etc. Modeling the quantized sensory input provides a fine-grained control over the amount of information required to solve the task at hand. This granularity coupled with optimal modeling (**Ch. 2**) can reduce evidence accumulation time while maintaining accuracy in many applications, such as (1) lowlight object recognition in **Ch. 4**, (2) modeling decision making processes in biological mechanisms in situations where both time and accuracy are important, e.g. **Ch. 3** and **Ch. 5**, and (3) preparing algorithms for next generation sensors that faithfully report the quantized signal.

Our analysis focuses on producing a correct decision as quickly as possible from quantized sensory inputs. We rely on the sequential probability ratio test (SPRT) for optimizing the speed versus accuracy tradeoff (SAT). Standard SPRT assumes that a probabilistic model is available to interpret the sensory inputs and that the model is constant over time (**Ch. 2**). We demonstrate three examples where these assumptions are satisfied to different extents. (1) In visual search (**Ch. 3**), both assumptions are satisfied, and SPRT is applied directly to account for ideal search performance and human behavior across different search environments. (2) In scotopic visual recognition (**Ch. 4**), the probabilistic model is constant in time but not available. This is common among practical applications that involve images, language and sound. We develop strategies to learn SPRT discriminately from data, and demonstrate that 1 photon per pixel is required for classifying black and white images of digits and about 20 are required for classifying color images of common objects (cats, dogs, cars, airplanes, etc). (3) In ecological situations such as visual discrimination with unknown onset (**Ch. 5**), the probabilistic model is known but not constant over time. We demonstrate methods to jointly infer the model and perform SPRT. We also discover that humans do not behave according to this model, but rather rely on a sub-optimal model with a simpler architecture.

In all applications, the quantized inputs are assumed to be Poisson in nature: pho-

tons that arrive at camera sensors and action potentials generated by orientation / motion-tuning neurons in earlier sensory systems both follow a homogeneous Poisson distribution. For Poisson distributed events (photons, action potentials), the sufficient statistics are the mean event rate, and the events are uncorrelated in time. It is therefore tempting to conclude that no algorithm can do better than the one that takes the mean event rate as input (which corresponds to the intensity image estimated from photon counts, and the neuron firing rate profile estimated from trains of action potentials). We demonstrate that this is *not* true, as this algorithm fails to consider the uncertainty associated with the mean estimates. For example, one may estimate a $10Hz$ firing rate from having observed two action potentials in $20ms$, but the $[10\%, 90\%]$ confidence region of the estimate is $[26, 200]Hz$, meaning that repeating the same observation may result in a rate estimate that is an order of magnitude larger. Therefore an algorithm that is aware of this uncertainty is likely to do better. Indeed, SPRT relies on the uncertainty to decide when a sufficient amount of evidence has been collected (**Ch. 2**), and empirically in the scotopic vision application **Sec. 4.4**, the WaldNet algorithm that incorporates the uncertainty outperforms the rate-based algorithm that does not. One ramification of the comparison is that images may not be the best medium for representing the visual world. This is because (1) images throw away the uncertainty information, and (2) in situations that demand fast and accurate decisions, acquisition time of the image may be undesirably long. Therefore, the computer vision community should not fixate on images, and instead start to consider photon streams, which are made available by recent sensor technologies [1]–[3].

Quantization occurs not just in the sensory inputs, but also on the internal computations of vision systems. We show that SPRT for visual search **Sec. 3.6** admits a spiking implementation. Log likelihoods of internal variables of the SPRT are represented as neurons that compute and communicate using action potentials. The computation is incremental: as quantized input comes in, only a sparse set of changes propagate through the network. The spiking implementation makes use of a small number of action potentials in total, and approximates SRPT well.

Many issues remain for future investigation. First, there lacks a hardware implementation that connects SPRT with photon counting sensors. The sensors may report photon counts at high spatial frequencies (e.g. a Single-Photon-Avalanche-Diode operates [1] at $10^9 Hz$), but current hardware implementations of convolutional networks are at the level of kHz [4]. While quantization of computation may be key

to further accelerate the system, there may also be an intermediate level of granularity between single photons and the high-quality image that makes sense for most lowlight vision applications.

Second, we have only explored learning algorithms (**Sec. 4.3**) for static models. In problems where the probabilistic models are unknown and non-static, one needs to simultaneously learn the model and apply optimal sequential testing accordingly. This is similar in the visual discrimination with unknown stimulus onset example **Ch. 5**, where the non-static model is parameterized by the stimulus onset, therefore SPRT addresses this issue by jointly estimating the onset timing and classifying the stimulus class. We are currently investigating scotopic tracking applications [5] where the dynamical model is fully parameterized by its initial conditions.

Lastly, active sensing may further improve the trade off between evidence accumulation cost and accuracy. We have so far assumed that the camera collects information passively for every pixel, whereas the camera could actively shut down pixels depending on their significance towards decision accuracy. The passive scheme makes sense when the goal is to minimize acquisition time, as we would like to maximize the amount of exposure for all pixels. However, if the goal is to minimize the total photon exposure, e.g. in biological imaging and surveillance applications, then it is reasonable to only collect from pixels that are most relevant to reach a decision. We speculate an algorithm that runs SPRT at every single pixel to determine the evidence accumulation time, in conjunction with the SPRT based on their outputs to compute the final decision. In either case, as we venture deep into the realm of quantized computation, the conventional notation of image becomes increasingly obsolete, and we should start to embrace the visual world as what it truly is – an ocean of photons. The image is just the waves that carry shells to the shore, the ocean is where the real treasures are.

References

[1] F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.

[2] L. Sbaiz, F. Yang, E. Charbon, S. Süsstrunk, and M. Vetterli, "The gigavision camera," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, 2009, pp. 1093–1096.

[3] E. Fossum, "The quanta image sensor (qis): Concepts and challenges," in *Imaging Systems and Applications*, Optical Society of America, 2011, JTuE1.

[4] K. Ovtcharov, O. Ruwase, J.-Y. Kim, J. Fowers, K. Strauss, and E. S. Chung, "Accelerating deep convolutional neural networks using specialized hardware," *Microsoft Research Whitepaper*, vol. 2, 2015.

[5] B. Chen and P. Perona, "Vision without the image," *Sensors*, vol. 16, no. 4, pp. 484–484, 2016.

1

Appendix A

APPENDIX

A.1 Visual search

Orientation log likelihood $\mathcal{L}_\theta$

We first derive how to compute the log likelihood for each task-relevant orientation from evidence $\boldsymbol{X}_{1:t}$ (in this section we are concerned with one location only, therefore we omit the location superscript l to simplify notation), which is a set of spike trains from N orientation-tuned neurons (which can be generalized to be sensitive to color, intensity, etc) collected during the time interval $[0, t\Delta]$. Let $\boldsymbol{X}^{(i)}_{1:t}$ be the set of spikes from neuron i in the time interval $[0, t\Delta]$, N^i_t the number of spikes from neuron i in $\boldsymbol{X}^i_{1:t}$, and N_t the total number of spikes, then the likelihood of $\boldsymbol{X}^{(i)}_{1:t}$ when stimulus orientation is θ is given by a Poisson distribution:

$$P(\boldsymbol{X}^{(i)}_{1:t}|Y=\theta) = \text{Poiss}(N^i_t|\lambda^i_\theta t) = (\lambda^i_\theta t)^{N^i_t}\frac{\exp(-\lambda^i_\theta t)}{N^i_t!}, \tag{A.1}$$

where λ^i_θ is the firing rate of neuron i when the stimulus orientation is θ.

The observations from the hypercolumn neurons are independent from each other, thus the log likelihood of $\boldsymbol{X}_{1:t}$ is given by:

$$\begin{aligned}\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) &\triangleq \log P(\boldsymbol{X}_{1:t}|Y=\theta) = \log\prod_{i=1}^{N} P(\boldsymbol{X}^{(i)}_{1:t}|Y=\theta)\\ &= \sum_{i=1}^{n_H}\log\left((\lambda^i_\theta t)^{n_H{}^i_t}\frac{\exp(-\lambda^i_\theta t)}{N^i_t!}\right)\\ &= \sum_{s=1}^{N_t} W^{i(s)}_\theta - t\sum_{i=1}^{n_H}\lambda^i_\theta + \text{const},\end{aligned} \tag{A.2}$$

where $W^i_\theta = \log\lambda^i_\theta$ is the contribution of each action potential from neuron i to the log likelihood of orientation θ, and "const" is a term that does not depend on θ and is therefore irrelevant for the decision. The first term is the "diffusion" that introduces jumps in $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ whenever a spike occurs. The second term is a "drift" term that moves $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ gradually in time. When the tuning curves of the neurons tessellate

regularly the circle of orientations, as is the case in our model (**Fig. 3.4a**), the average firing rate of the hypercolumn under different orientations is approximately the same, and the drift term may be safely omitted from models.

Review: Bayesian inference for discrimination and homogeneous search

We first re-derive the log likelihood ratio $S(\boldsymbol{X}_{1:t})$ for visual discrimination. For all derivations below we show how to compute $\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)}$ from the orientation log likelihoods $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$, keeping in mind that

$$S(\boldsymbol{X}_{1:t}) = \log \frac{P(C=1|\boldsymbol{X}_{1:t})}{P(C=0|\boldsymbol{X}_{1:t})} = \log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} + \log \frac{P(C=1)}{P(C=0)}.$$

In homogeneous discrimination, the target and distractor have distinct and unique orientations θ_T and θ_D, therefore:

$$\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} = \log \frac{P(\boldsymbol{X}_{1:t}|\theta=\theta_T)}{P(\boldsymbol{X}_{1:t}|\theta=\theta_D)} = \mathcal{L}_{\theta_T}(\boldsymbol{X}_{1:t}) - \mathcal{L}_{\theta_D}(\boldsymbol{X}_{1:t}), \tag{A.3}$$

which proves **Eq. 3.3**.

In heterogeneous discrimination, $\theta_T \in \Theta_T$ and $\theta_D \in \Theta_D$. For simplicity assume uniform prior on both target and distractor orientation, i.e. $P(\theta|C=1) = 1/n_T, \forall\theta \in \Theta_T$ and $P(\theta|C=0) = 1/n_D, \forall\theta \in \Theta_D$:

$$\begin{aligned}
\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} &= \log \frac{P(\boldsymbol{X}_{1:t}|\theta\in\Theta_T)}{P(\boldsymbol{X}_{1:t}|\theta\in\Theta_D)} \\
&= \log\left(\sum_{\theta\in\Theta_T} P(\boldsymbol{X}_{1:t}|\theta)P(\theta|C=1)\right) - \log\left(\sum_{\theta\in\Theta_D} P(\boldsymbol{X}_{1:t}|\theta)P(\theta|C=0)\right) \\
&= \log\left(\sum_{\theta\in\Theta_T} \frac{\exp(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}))}{n_T}\right) - \log\left(\sum_{\theta\in\Theta_D} \frac{\exp(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}))}{n_D}\right) \\
&= \underset{\theta\in\Theta_T}{\mathcal{S}\max}\left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) - \log(n_T)\right) - \underset{\theta\in\Theta_D}{\mathcal{S}\max}\left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}) - \log(n_D)\right),
\end{aligned} \tag{A.4}$$

which proves **Eq. 3.5**.

Now we re-derive $S(\boldsymbol{X}_{1:t})$ for homogeneous visual search ($M = L > 1, n_T = n_D = 1$) from the local orientation log likelihoods $\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t})$ from each of the L locations.

Call $l_T \in \{1, 2, \ldots, L\}$ the target location and assume uniform prior on l_T. **Eq. 3.3** is proved below:

$$\begin{aligned}\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} &= \log \frac{\sum_{l_T} P(\boldsymbol{X}_{1:t}|l_T)P(l_T|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} = \log \frac{1}{L}\sum_{l_T} \frac{P(\boldsymbol{X}_{1:t}|l_T)}{P(\boldsymbol{X}_{1:t}|C=0)} \\ &= \log \frac{1}{L}\sum_{l_T} \frac{P(\boldsymbol{X}^{l_T}_{1:t}|\theta_T)\prod_{l\neq l_T} P(\boldsymbol{X}^{l}_{1:t}|\theta_D)}{\prod_{l} P(\boldsymbol{X}^{l}_{1:t}|\theta_D)} \\ &= \log \frac{1}{L}\sum_{l_T} \frac{P(\boldsymbol{X}^{l_T}_{1:t}|\theta_T)}{P(\boldsymbol{X}^{l_T}_{1:t}|\theta_D)} = \mathcal{S}\max_{l_T}\left(\mathcal{L}_{\theta_T}(\boldsymbol{X}^{l_T}_{1:t}) - \mathcal{L}_{\theta_D}(\boldsymbol{X}^{l_D}_{1:t}) - \log(L)\right).\end{aligned} \tag{A.5}$$

Formulating common search problems using the general model

The heterogeneous visual search model is a general model for explaining a wide range of search tasks. The general model captures the variability in set-size and orientation contrast using CDD, which is the distribution $P(Y^l|C^l = 0)$ of stimulus orientation at a non-target location. Below are three examples:

Mixed contrast (**Exp. 2**): the distractor orientation is sampled uniformly from $\{20°, 30°, 45°\}$, and all the distractors must have the same orientation.

In this case a CDD is a three dimensional vector of

$$\phi = [P(Y^l = 20°|C^l = 0), P(Y^l = 30°|C^l = 0), P(Y^l = 45°|C^l = 0)].$$

We will employ three CDDs:

$$\phi^{(1)} = [1, 0, 0]; \phi^{(2)} = [0, 1, 0]; \phi^{(3)} = [0, 0, 1];$$

with equal prior probability $P(\phi^{(i)}) = 1/3, \forall i$.

This setup exactly describes the probabilistic structure of **Exp. 2**. Since each CDD is a delta function at a single orientation, distractors at all locations will be identical, and the distractor orientations will be chosen uniformly at random from $\{20°, 30°, 45°\}$.

I.i.d. distractor heterogeneous search: the distractor orientation is sampled independently at each location from $\{20°, 30°, 45°\}$ with probability $[0.2, 0.5, 0.3]$.

This is precisely the i.i.d. distractor heterogeneous search task (**Eq. 3.9**). Only one CDD is needed, and $\phi = [0.2, 0.5, 0.3]$.

Mixed set-size (**Exp. 3**): the distractor orientation is 30°. The set-size M is sampled uniformly from $\{3, 6, 12\}$. The total number of display locations is $L = 12$.

In this case, denote $Y^l = \emptyset$ that a non-target location is blank. If there are M display items, then the probability of any non-target location being blank is $(L - M)/L$. A CDD is a two dimensional vector of

$$\phi = [P(Y^l = 20°|C^l = 0), P(Y^l = \emptyset|C^l = 0)],$$

and the three different set-sizes may be represented by three CDDs of equal probability:

$$\phi^{(1)} = [3/12, 9/12], \phi^{(2)} = [6/12, 6/12], \phi^{(3)} = [1 - \epsilon, \epsilon], \tag{A.6}$$

where ϵ is a small number to prevent zero probability.

> Note that the setup in **Eq. A.6** only approximates the probabilistic structure of **Exp. 3**. This is because the blank placements are not independent of one another. In other words, for a given set-size M, only M locations can contain a distractor. If we place a distractor at each location with probability M/L, we do not always observe M distractors. Instead, the actual set-size follows a binomial distribution with mean M. However, this is a reasonable approximation because the human visual system can generalize to unseen set-sizes effortlessly. In addition, the values of M used in our experiments are often different enough $\{3, 6, 12\}$ that the i.i.d. model is equally effective in inferring M.

Bayesian inference for heterogeneous visual search

SPRT relies on computing $S(\boldsymbol{X}_{1:t})$ from the orientation log likelihoods $\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t})$ from all locations l, which we show below. The target-present likelihood $P(\boldsymbol{X}_{1:t}|C = 1)$ is given by marginalizing out the target location $l_T \in \{1, 2, \ldots, L\}$, CDD ϕ, as well as the target and distractor orientations. Denote $C^l \in \{0, 1\}$ the stimulus class at location l: $C^l = 1$ if and only if location l contains a target. In light of the graphical model in **Fig. 3.1b**:

$$
\begin{aligned}
P(\boldsymbol{X}_{1:t}|C=1) &= \sum_{l_T,\phi} P(\boldsymbol{X}_{1:t}|l_T,\phi,C=1)P(\phi)P(l_T|C=1) \\
&= \sum_{l_T} P(l_T|C=1)\sum_{\phi} P(\phi) \sum_{\vec{Y}=\{Y^1,\ldots,Y^L\}} P(\boldsymbol{X}_{1:t}|\vec{Y})P(\vec{Y}|l_T,\phi,C=1) \\
&= \sum_{l_T} P(l_T|C=1)\sum_{\phi} P(\phi) \sum_{\vec{Y}} \prod_{l} (P(\boldsymbol{X}^l_{1:t}|Y^l)P(Y^l|l_T,\phi,C=1)) \\
&= \sum_{l_T} P(l_T|C=1)\sum_{\phi} P(\phi) \prod_{l} \sum_{Y^l} (P(\boldsymbol{X}^l_{1:t}|Y^l)P(Y^l|l_T,\phi,C=1)) \\
&= \sum_{l_T} P(l_T|C=1)\sum_{\phi} P(\phi)P(\boldsymbol{X}^{l_T}_{1:t}|C^{l_T}=1) \prod_{l\neq l} P(\boldsymbol{X}^l_{1:t}|\phi,C^l=0) \\
&= \sum_{l_T} P(l_T|C=1)\sum_{\phi} \frac{P(\boldsymbol{X}^{l_T}_{1:t}|C^{l_T}=1)}{P(\boldsymbol{X}^{l_T}_{1:t}|\phi,C^{l_T}=0)} P(\phi) \prod_{l} P(\boldsymbol{X}^l_{1:t}|\phi,C^l=0),
\end{aligned}
\tag{A.7}
$$

where

$$
\begin{aligned}
P(\boldsymbol{X}^l_{1:t}|C^l=1) &= \sum_{\theta\in\Theta_T} P(\boldsymbol{X}^l_{1:t}|Y^l=\theta)P(\theta|C^l=1), \\
P(\boldsymbol{X}^l_{1:t}|\phi,C^l=0) &= \sum_{\theta\in\Theta_D} P(\boldsymbol{X}^l_{1:t}|Y^l=\theta)\phi_\theta.
\end{aligned}
$$

Similarly, the target-absent likelihood is:

$$
P(\boldsymbol{X}_{1:t}|C=0) = \sum_{\phi} P(\phi) \prod_{l} P(\boldsymbol{X}^l_{1:t}|\phi,C_l=0). \tag{A.8}
$$

Note that **Eq. A.8** may be thought of as computing a normalization of the term $P(\phi)\prod_l P(\boldsymbol{X}^l_{1:t}|\phi,C^l=0)$ that is used to weight the local log likelihood ratios in **Eq. A.7**. This normalized weight turns out to be the posterior of CDD: $P(\phi|\boldsymbol{X}_{1:t})$. Define the log posterior of CDD as:

$$
Q_\phi(\boldsymbol{X}_{1:t}) \triangleq \log P(\phi|\boldsymbol{X}_{1:t}) = \log \frac{P(\phi)\prod_l P(\boldsymbol{X}^l_{1:t}|\phi,C^l=0)}{\sum_{\phi'} P(\phi')\prod_l P(\boldsymbol{X}^l_{1:t}|\phi',C^l=0)}. \tag{A.9}
$$

Then the log likelihood ratio is

$$\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} = \log \sum_l P(l_T = l|C=1) P(\boldsymbol{X}^l_{1:t}|C^l=1) \sum_\phi \frac{P(\phi|\boldsymbol{X}_{1:t})}{P(\boldsymbol{X}^l_{1:t}|\phi, C^l=0)}.$$

$$\text{Recall that: } \underset{i\in A}{\mathcal{S}\max}\,(x_i) = \log \sum_{i\in A} \exp(x_i), \tag{A.10}$$

$$\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)}$$
$$= \underset{l=1,\ldots,L}{\mathcal{S}\max} \left(\log P(l_T = l|C=1) + \log P(\boldsymbol{X}^l_{1:t}|C^l=1) + \underset{\phi\in\Phi}{\mathcal{S}\max} \left(Q_\phi(\boldsymbol{X}_{1:t}) - \log P(\boldsymbol{X}^l_{1:t}|\phi, C^l=0) \right) \right). \tag{A.11}$$

Assuming uniform prior on the target location $P(l_T = l|C=1)$ and on the target type $P(Y^l = \theta|C^l=1)$,

$$\begin{aligned} \log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} =& \underset{l=1\ldots,L}{\mathcal{S}\max}\,(A+B) - \log(L), \\ \text{where } A =& \underset{\theta\in\Theta_T}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) - \log(n_T) \right), \\ B =& \underset{\phi\in\Phi}{\mathcal{S}\max} \left(-\underset{\theta\in\Theta_D}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) + \log\phi_\theta \right) + Q_\phi(\boldsymbol{X}_{1:t}) \right), \end{aligned} \tag{A.12}$$

which proves **Eq. 3.10-3.11**.

Mean-field approximation to SPRT

Instead of inferring the CDD on a trial-by-trial basis, a simpler alternative is to use its average value without looking at the stimulus. For example, in the mixed set-size example with $M \in \{3, 6, 12\}$, SPRT estimates the value of M given $\boldsymbol{X}_{1:t}$ for each trial, whereas the simple model assumes a set-size of $\mathbb{E}(M) = 7$ for all the trials.

In detail, the simple model essentially uses the 'mean-field' approximation on **Eq. A.12**:

$$\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} \approx \underset{l=1,\ldots,L}{\mathcal{S}\max} \left(\underset{\theta\in\Theta_T}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) \right) - \underset{\theta\in\Theta_D}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}^l_{1:t}) + \log\bar{\phi}_\theta \right) \right) - \log(n_T L), \tag{A.13}$$

where $\bar{\phi}_\theta = \sum_{\phi\in\Phi} \phi_\theta P(\phi)$ is the mean CDD with respect to the its prior distribution. The prediction of the simple model on a mixed-set-size search problem is shown in **Fig. 3.8b**.

Search with correlated target and distractor orientations

SPRT for heterogeneous visual search **Eq. A.12** assumes that the properties of the scene, namely the set-size and the scene-complexity, only affects the distractor orientation distribution. In this section we relax this assumption and let ϕ encode both the target and distractor orientation distribution: $\phi = \{\phi^{(T)}, \phi^{(D)}\}$, where $\phi_\theta^{(T)} = P(Y^l = \theta | C^l = 1)$ and $\phi_\theta^{(D)} = P(Y^l = \theta | C^l = 0)$. The log likelihood of target-present in **Eq. A.7** now becomes:

$$P(\boldsymbol{X}_{1:t}|C=1) = \sum_{l_T} P(l_T|C=1) \sum_{\phi} \frac{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(T)}, C^{l_T}=1)}{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi) \prod_l P(\boldsymbol{X}_{1:t}^l|\phi^{(D)}, C^l=0).$$

The log likelihood ratio of target-present versus target-absent is:

$$\begin{aligned}
\log \frac{P(\boldsymbol{X}_{1:t}|C=1)}{P(\boldsymbol{X}_{1:t}|C=0)} &= \log \sum_l P(l_T=l|C=1) \sum_{\phi} \frac{P(\boldsymbol{X}_{1:t}^l|\phi^{(T)}, C^l=1)}{P(\boldsymbol{X}_{1:t}^l|\phi^{(D)}, C^l=0)} P(\phi|\boldsymbol{X}_{1:t}) \\
&= \underset{l=1,\ldots,L}{\mathcal{S}\max} \left(\underset{\phi\in\Phi}{\mathcal{S}\max} \left(A(l,\phi) \right) \right) - \log(L), \\
\text{where } A(l,\phi) &= \underset{\theta\in\Theta_T}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}^l) + \log \phi_\theta^{(T)} \right) - \underset{\theta\in\Theta_D}{\mathcal{S}\max} \left(\mathcal{L}_\theta(\boldsymbol{X}_{1:t}^l) + \log \phi_\theta^{(D)} \right) + Q_\phi(\boldsymbol{X}_{1:t}).
\end{aligned} \tag{A.14}$$

This formulation encompasses the formulation in **Eq. A.12** where the target and the distractor orientations are distributed independently with respect to each other. To see this, assume $\phi^{(D)}$ and $\phi^{(T)}$ vary independently, then:

$$\begin{aligned}
&P(\boldsymbol{X}_{1:t}|C=1) \\
&= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(T)},\phi^{(D)}} \frac{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(T)}, C^{l_T}=1)}{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(T)})P(\phi^{(D)}) \prod_l P(\boldsymbol{X}_{1:t}^l|\phi^{(D)}, C^l=0) \\
&= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(D)}} \frac{\sum_{\phi^{(T)}} P(\phi^{(T)}) P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(T)}, C^{l_T}=1)}{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(D)}) \prod_l P(\boldsymbol{X}_{1:t}^l|\phi^{(D)}, C^l=0) \\
&= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(D)}} \frac{P(\boldsymbol{X}_{1:t}^{l_T}|\bar{\phi}^{(T)}, C^{l_T}=1)}{P(\boldsymbol{X}_{1:t}^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(D)}) \prod_l P(\boldsymbol{X}_{1:t}^l|\phi^{(D)}, C^l=0),
\end{aligned} \tag{A.15}$$

where $\bar{\phi}^{(T)} = \sum_{\phi^{(T)}} \phi^{(T)} P(\phi^{(T)})$ is the expected value of $\phi^{(T)}$. **Eq. A.15** is equivalent to **Eq. A.12** with a different prior ($\bar{\phi}^{(T)}$) on target orientation.

A.2 Scotopic visual recognition

Time-adaptation of hidden features (Eq. 4.8)

We explain how to compute hidden features $S^H(\boldsymbol{N}_t)$ from partial observations $\boldsymbol{N}_t$, where $t \leq T$ and T is the exposure time required to obtain a high-quality image. In order to compute $\boldsymbol{h}$ we need to marginalize out the unobserved photons $\Delta\boldsymbol{N} = \sum_{t'=t+1}^{T} \boldsymbol{X}_{t'}$:

$$S_j^H(\boldsymbol{N}_t) = \sum_{\Delta\boldsymbol{N}} S_j^H(W_j(\boldsymbol{N}_t + \Delta\boldsymbol{N}) + b_j)P(\Delta\boldsymbol{N}|\boldsymbol{N}_t). \tag{A.16}$$

To approximate the marginalization above, we put a Gamma prior on the photon emission rate λ_i at pixel i:

$$P(\lambda_i) = Gam(\mu_i\sigma_\lambda, \sigma_\lambda). \tag{A.17}$$

After observing the cumulative count $N_{i,t}$ of pixel i in time $[0, t\Delta]$, the posterior estimate for the photon emission rate is:

$$P(\lambda_i|N_{i,t}) \propto P(N_{i,t}|\lambda_i)P(\lambda_i) \tag{A.18}$$

$$= Gam(\mu_i\sigma_\lambda + N_{i,t}, \sigma_\lambda + t), \tag{A.19}$$

which has a posterior mean of:

$$\hat{\lambda}_i \triangleq \mathbb{E}[\lambda_i|N_{i,t}] = \frac{\mu_i\sigma_\lambda + N_{i,t}}{\sigma_\lambda + t}. \tag{A.20}$$

Intuitively, the emission rate is estimated via a smoothed-average of the observed counts.

Therefore the marginalization step in **Eq. A.16** may be approximated up to second order accuracy using:

$$P(h_j = 1|\boldsymbol{N}_t) \approx P(h_j = 1|\mathbb{E}[\Delta\boldsymbol{N}|\boldsymbol{N}_t] + \boldsymbol{N}_t) \tag{A.21}$$

$$= Sigm(\sum_i W_{ji}\left((T-t)\hat{\lambda}_i + X_{i,t}\right) + b_j) \tag{A.22}$$

$$= Sigm(W_j\frac{T+\sigma_\lambda}{t+\sigma_\lambda}\boldsymbol{N}_t + \sigma_\lambda\frac{T-t}{\sigma_\lambda+t}W_j\mu + b_j) \tag{A.23}$$

$$= Sigm(\alpha(t)W_j\boldsymbol{N}_t + \beta_j(t)), \tag{A.24}$$

$$\text{where} \quad \alpha(t) \triangleq \frac{T+\sigma_\lambda}{t+\sigma_\lambda}, \tag{A.25}$$

$$\beta_j(t) \triangleq \sigma_\lambda\frac{T-t}{\sigma_\lambda+t}W_j\mu + b_j, \tag{A.26}$$

thus the log posterior ratio is:

$$S_j^H(\boldsymbol{N}_t) = \log \frac{P(h_j = 1|\boldsymbol{N}_t)}{P(h_j = 0|\boldsymbol{N}_t)} \approx \alpha(t) W_j \boldsymbol{N}_t + \beta_j(t), \tag{A.27}$$

which proves **Eq. 4.8**.

The derivation above was done for the W_j-th feature only. In ConvNet, the features are localized (e.g. occupying only a 5×5 region), and organized into groups (e.g. the first layer in WaldNet for CIFAR10 uses 32 features groups), which means that we need to learn one $\beta_j(t)$ for each spatial location and each feature group. For simplicity we assume that the mean image μ is translational invariant within 5×5 regions, so that we only need to model one scalar $\beta_j(t)$ for each of the 32 feature maps.

Relationship between exposure time and number of bits of signal (Table 4.1)
Bits of signal and photon counts are equivalent concepts. Furthermore, the photon counts are linearly related to exposure time. Here we derive the relationship between exposure time and the number of bits of signal. To simplify the analysis we will make the assumption that our imaging setup has a constant aperture.

What does it mean for an image to have a given number of bits of signal? Each pixel is a random variable reproducing the brightness of a piece of the scene up to some noise. There are two main sources of noise: the electronics and the quantum nature of light. We will assume that for bright pixels the main source of noise is light. This is because, as will be clear from our experiments, a fairly small number of bits per pixel are needed for visual classification, and current image sensors and AD converters are more accurate than that.

According to the Poisson noise model (**Eq. 4.4** in main text), each pixel receives photons at rate λ. The expected number of photons collected during a time t is λt and the standard deviation is $\sigma = \sqrt{\lambda t}$. We will ignore the issue of quantum efficiency (QE), i.e. the conversion rate from photons to electrons on the pixel's capacitor, and assume that QE=1 to simplify the notation (real QEs may range from 0.5 to 0.8). Thus, the SNR of a pixel is $SNR = \lambda t / \sqrt{\lambda t} = \sqrt{\lambda t}$ and the number of bits of signal is $b = \log_2 \sqrt{\lambda t} = 0.5 \log_2 \lambda + 0.5 \log_2 t$.

The value of λ depends on the amount of light that is present. This may change dramatically: from 10^{-3} LUX in a moonless night to 10^5 LUX in bright direct sunlight. With a typical camera one may obtain a good quality image in a well lit

indoor scene ($E_v \approx 300$ lux) with an exposure time of 1/30s. If a bright pixel has 6.5 bits of signal, the noise is $2^{-6.5} \approx 1\%$ of the dynamic range and $\lambda t/\sqrt{\lambda t} = 100$, i.e. $\lambda \approx 3 \cdot 10^5 \approx 10^3 E_v \approx 2^{10} E_v$. Substituting this calculation of λ into the expression derived in the previous paragraph we obtain $b \approx 5 + \frac{1}{2}\log_2 t + \frac{1}{2}\log_2 E_v$, which is what we used to generate **Table 4.1** in the main text.

Datasets

MNIST contains gray-scaled 28×28 images of 10 hand-written digits. It has $50k$ training and $10k$ test images. We treat the pixel values as the ground truth intensity[1]. Dark current $\epsilon_{dc} = 3\%$. We use the default 'LeNet' from the MatConvNet package [1]. The architecture is 784-20-50-500-10[2] with 5×5 receptive fields and 2×2 pooling.

CIFAR10 contains 32×32 color images of 10 visual categories. It has $50k$ training and $10k$ test images. We use the same synthesis procedure as above for each color channel[3]. We use the default 1024-32-32-64-10 LeNet architecture [2] with batch normalization [3] after each convolution layer. We use the same setting prescribed in [2] to achieve 18% test error on normal lighting conditions. [2] uses local contrast normalization and ZCA whitening as preprocessing steps. We estimate the local contrast and ZCA from normal lighting images and transform them according to the lowlight model to preprocess scotopic images. We use batch-normalization to accelerate learning. All models are trained for 75 epochs, where the learning rate is 0.05 for 30 iterations, 0.005 for the next 25 then 0.0005 for the rest.

Implementation

In step one of learning, the scalar functions $\alpha(t)$ and $\boldsymbol{\beta}(t)$ in **Eq. 4.8** are learned as follows. As the inputs to the network are preprocessed, the preprocessing steps alter the algebraic form for α and $\boldsymbol{\beta}$. For flexibility we do not impose parametric forms on α and $\boldsymbol{\beta}$, but represent them with piecewise cubic Hermite interpolating polynomials with four end points at PPP= $[.22, 2.2, 22, 220]$. We learned the adapted weights at these end-points by using a different batch normalization module for each PPP. At

[1] The brightest image we synthesize has about 2^8 photons, which corresponds to a pixel-wise maximum signal-to-noise ratio of 16 (4-bit accuracy), whereas the original MNIST images has an accuracy of 7 to 8 bits, which corresponds to $2^{14} \sim 2^{16}$ photons.

[2] The first and last number represent the input and output dimension, each number in between represents the number of feature maps used for that layer. The number of units is the product of the number of feature maps with the size of the input.

[3] For simplicity we do not model the Bayer filter mosaic.

test time the parameters of the modules are interpolated to accommodate other PPP levels.

In step two of learning, we compute $S_c(\boldsymbol{N}_t)$ for 50 uniformly spaced PPPs in log scale, and train thresholds $\tau(t)$ for each PPP and for each η. A regularizer $0.01 \sum_t ||\tau(t) - \tau(t+1)||^2$ is imposed on the thresholds $\tau(t)$ on the log posterior ratios to enforce smoothness. In **Eq. 4.14**, the steepness of Sigmoid $Sigm$ is annealed over 500 iterations of gradient descent, with initial value 0.5, a decay rate of 0.99 and a floor value of 0.01.

A.3 Visual discrimination with unknown stimulus onset

Log posterior ratios based on momentary observations

Consider a visual display at time interval of motion coherence z over the time interval $[t\Delta, (t+1)\Delta]$. We make the simplifying assumption that each dot has probably z of moving along the coherent direction that is *independent* of the motion direction of the other dots. This means that there will be zM dots moving coherently *on average*, but at any point time, the actual number of coherently moving dots follows a multinomial distribution centered at zM. This is ok because the visual system should still function when it sees a slightly different number of moving dots than the expected value.

Let Y and $Y_i \in [0°, 360°]$ denote, respectively, the direction of coherent motion and the direction of local motion at location i. Z denotes the coherence. X_i is the instantaneous firing pattern of all locations and all hypercolumn neurons, and $X_{i,t}$ the pattern for location i. The likelihood of observing a firing pattern X of dots moving towards direction θ at coherence level z is:

$$P(\boldsymbol{X}|Y=\theta, Z=z) = \sum_{Y_1,Y_2,\ldots,Y_M} P(X|Y_1,Y_2,\ldots,Y_M)P(Y_1,Y_2,\ldots,Y_M|Z=z,Y=\theta) \tag{A.28}$$

$$= \sum_{Y_1,Y_2,\ldots,Y_M} \left(\prod_i P(X_i|Y_i)\right)\left(\prod_i P(Y_i|Z=z,Y=\theta)\right) \tag{A.29}$$

$$= \prod_i \sum_{Y_i} P(X_i|Y_i)P(Y_i|Z=z,Y=\theta). \tag{A.30}$$

Note that **Eq. A.29** makes critical use of the independence assumption of motion directions across locations, without which the computation would be intractable. Consider the term $P(Y_i|z, Y=\theta)$: the local direction Y_i should be θ if the dot is in

the coherent set and sampled uniformly from $[0°, 360°]$ otherwise, thus:

$$P(Y_i|Z = z, Y = \theta) = z^{\mathbb{I}[Y_i=\theta]} \left((1-z)\text{Uniform}(Y_i|[-180, 180])\right)^{\mathbb{I}[Y_i\neq\theta]}. \quad \text{(A.31)}$$

Making use of the fact that for all the incoherent directions the local direction prior is identical:

$$\begin{aligned} P(\boldsymbol{X}|Z = z, Y = \theta) &= \prod_i \sum_{Y_i} P(X_i|Y_i)P(Y_i|p, Y = \theta) \\ &= \prod_i \left((1-z)\mathbb{E}_{Y_i}[P(X_i|Y_i)] + zP(X_i|Y_i = \theta)\right). \end{aligned} \quad \text{(A.32)}$$

Therefore, the log likelihood ratio $r^{1,0} \triangleq \log \frac{P(X|Y=\theta,Z=z)}{P(X|Z=0)}$ between coherence z and coherence 0 is given by:

$$r^{1,0} = \sum_i S_i(X_i) \quad \text{(A.33)}$$

$$\text{where } S_i(X_i) \triangleq \log \frac{\left((1-z)\mathbb{E}_{Y_i}[P(X_i|Y_i)] + zP(X_i|Y_i = \theta)\right)}{\mathbb{E}_{Y_i}[P(X_i|Y_i)}. \quad \text{(A.34)}$$

When Δ is sufficiently short (say $< 1ms$) we can assume that there is at most one action potential in each hypercolumn. Let $I(X_i) \in \{0, \ldots, K\}$ denote the index of the firing neuron at location i. $I(X_i) = 0$ means there are no spikes. Here $I(X_i)$ and X_i are two representations of the same variable. According to **Eq. 5.7**, the probability of observing a spike from neuron k is $P(I(X_i) = k|Y_i = \theta) = \lambda_k^\theta \Delta$, and the probability for no spike is: $P(I(X_i) = 0|Y_i = \theta) = 1 - \sum_k \lambda_k^\theta \Delta$. We have for $k > 0$:

$$W_k^{1,0} \triangleq S_i(X_i : I(X_i) = k) = \log \frac{\left((1-z)\mathbb{E}_{Y_i}[\lambda_k^{Y_i}\Delta] + z\lambda_k^\theta\Delta\right)}{\mathbb{E}_{Y_i}\lambda_k^{Y_i}\Delta} = \log \frac{(1-z)\bar{\lambda} + z\lambda_k^\theta}{\bar{\lambda}}, \quad \text{(A.35)}$$

where $\bar{\lambda} \triangleq \mathbb{E}_d[\lambda_k^\theta]$ is a neuron's average firing rate over all directions. Since this average rate is identical across neurons, $\bar{\lambda}$ does not have a neuron index. In the same fashion, $W_k^{1,0} \triangleq S_i(I(X_i) = k)$ does not have a location index.

When $k = 0$, we have:

$$W_0^{1,0} \triangleq S_i(X_i : I(X_i) = 0) = \log \frac{(1-z)\mathbb{E}_{Y_i}[1 - \sum_k \lambda_k^{Y_i}\Delta] + z(1 - \sum_k \lambda_k^\theta\Delta)}{\mathbb{E}_{Y_i}[1 - \sum_k \lambda_k^{Y_i}\Delta]} = 0. \quad \text{(A.36)}$$

Putting **Eq. A.35** and **Eq. A.36** together we have proven **Eq. 5.9**:

$$r^{1,0} = \sum_i W^{1,0}_{I(X_i)} = \sum_i \sum_k W^{1,0}_k X(i,k). \tag{A.37}$$

Similar derivations on $r^{1,2} \triangleq \log \frac{P(\boldsymbol{X}|D=\theta_1, Z=z)}{P(\boldsymbol{X}|Z=\theta_2, Z=z)}$ proves **Eq. 5.12**.

Log posterior ratios based on spike trains

Now we discuss how to compute $S^{c,0}_t \triangleq \log \frac{P(C_t=c|\boldsymbol{X}_{1:t})}{P(C_t=0|\boldsymbol{X}_{1:t})}$ based on observations from the entire duration of $[0, t\Delta]$. For now let us assume that there is only one coherent motion class c. We can compute the enumerator by marginalization over the change point t_δ:

$$P(C_t = c|\boldsymbol{X}_{1:t}) = \sum_{t_d=1}^{t} P(t_\delta = t_d|\boldsymbol{X}_{1:t}) \tag{A.38}$$

$$= \sum_{t_d=1}^{t} P(\boldsymbol{X}_{1:t}|t_\delta = t_d) P(t_\delta = t_d)/P(\boldsymbol{X}_{1:t}) \tag{A.39}$$

$$= \sum_{t_d=1}^{t} \left(\prod_{i=1}^{t_d-1} P(X_i|C_i = 0) \right) \left(\prod_{j=t_d}^{T} P(X_j|C_j = c) \right) P(t_\delta = t_d)/P(\boldsymbol{X}_{1:t}). \tag{A.40}$$

Similarly,

$$P(t_\delta = 0|\boldsymbol{X}_{1:t}) = \left(\prod_{i=1}^{t} P(X_i|C_i = 0) \right) P(t_\delta > t)/P(\boldsymbol{X}_{1:t}). \tag{A.41}$$

Taking the ratio between **Eq. A.40** and **Eq. A.41** gives:

$$S^{c,0}_t = \log \frac{P(C_t = c|\boldsymbol{X}_{1:t})}{P(C_t = 0|\boldsymbol{X}_{1:t})} = \log \left(\sum_{t_d=1}^{t} \left(\prod_{i=t_d}^{t} \frac{P(X_i|C_i = c)}{P(X_i|C_i = 0)} \right) \frac{P(t_\delta = t_d)}{P(t_\delta > t)} \right), \tag{A.42}$$

which admits the following recursive computation:

$$S^{c,0}_t = \log \left(\sum_{t_d=1}^{t-1} \left(\prod_{i=t_d}^{t-1} \frac{P(X_i|C_i = c)}{P(X_i|C_i = 0)} \right) \frac{P(X_t|C_t = c)}{P(X_t|C_t = 0)} \frac{P(t_\delta = t_d)}{P(t_\delta > t-1)} \frac{P(t_\delta > t-1)}{P(t_\delta > t)} + \frac{P(X_t|C_t = 1)}{P(X_t|C_t = 0)} \frac{P(t_\delta = t)}{P(t_\delta > t)} \right) \tag{A.43}$$

$$= \log \left(\left(\exp(S_{t-1}) \frac{P(t_\delta > t-1)}{P(t_\delta > t)} + \frac{P(t_\delta = t)}{P(t_\delta > t)} \right) \frac{P(X_t|C_t = 1)}{P(X_t|C_t = 0)} \right) \tag{A.44}$$

$$= \log \left(\exp \left(S_{t-1} - \log \frac{P(t_\delta = t)}{P(t_\delta > t-1)} \right) + 1 \right) + \log \frac{P(t_\delta = t)}{P(t_\delta > t)} + r^{c,0}_t \tag{A.45}$$

$$= \mathcal{S}\mathrm{rec} \left(S_{t-1} - \log \alpha_t \right) + \log \frac{\alpha_t}{1-\alpha_t} + r^{c,0}_t, \tag{A.46}$$

which, recalling that $\alpha_t \triangleq P(\kappa = t|\kappa > t-1)$, proves **Eq. 5.10**.

To relax the unique coherent motion assumption, one can simplify offset $S_t^{c,0}$ by the log prior $\log P(C=c)$ for the class c. To compute ratios between the two coherent motions (**Eq. 5.1**):

$$S_t^{i,j} \triangleq \log \frac{P(C_t = i|\boldsymbol{X}_{1:t})}{P(C_t = j|\boldsymbol{X}_{1:t})} = \log \left(\frac{P(C_t = i|\boldsymbol{X}_{1:t})}{P(C_t = 0|\boldsymbol{X}_{1:t})} / \frac{P(C_t = j|\boldsymbol{X}_{1:t})}{P(C_t = 0|\boldsymbol{X}_{1:t})} \right) = S_t^{i,0} - S_t^{j,0}. \tag{A.47}$$

Lastly, $R_{t,t'}$ (**Eq. 5.2**) the log posterior ratios for post-change observations is simply:

$$R_{t,t'} \triangleq \log \frac{P(C'_t = 1|X_{t:t'}, \kappa \le t}{P(C'_t = 2|X_{t:t'}, \kappa \le t} = \log \frac{P(C=1)}{P(C=2)} + \sum_i \log \frac{P(X_i|C_i=1)}{P(X_i|C_i=2)} = \log \frac{P(C=1)}{P(C=2)} + \sum_{i=t}^{t'} r_i^{1,2}, \tag{A.48}$$

which proves **Eq. 5.13**.

A.4 Optimality analysis

State formulation in visual search

We have chosen the log posterior ratios at all locations: $\vec{Z} : Z_l(t) = \log \frac{P(\boldsymbol{X}^l_{1:t}|C^l=1)}{P(\boldsymbol{X}^l_{1:t}|C^l=0)}$, $l = 1 \ldots M$, to be the state of our model because the resultant system is Markov: i.e. $\vec{Z}$ is a sufficient statistic to compute both the overall log likelihood ratio $S_{\text{homo-search}}$ and likelihood of future observations.

First, as shown in [4], [5]

$$S(\boldsymbol{X}_{1:t}) = \log \frac{P(C=1|\boldsymbol{X}_{1:t})}{P(C=0|\boldsymbol{X}_{1:t})} = \underset{l=1\ldots M}{\mathcal{S}\max}(Z_l) - \log(M).$$

Second, the likelihood of new observation $\boldsymbol{X}_{t+1}$ at time $t+1$ is obtained by marginalizing the target location l_T. Denote $l_T = 0$ the target-absent event:

$$P(l_T = 0|\boldsymbol{X}_{1:t}) = P(C=0|\boldsymbol{X}_{1:t}) = \frac{1}{1+\exp(S(\boldsymbol{X}_{1:t}))} = \frac{1}{1+\sum_l \exp(Z_l)/M},$$
$$P(l_T, l_T > 0|\boldsymbol{X}_{1:t}) = \frac{\exp(Z_l(t))/M}{1+\sum_l \exp(Z_l(t))/M}.$$

For notational convenience, define $Z_0 = \log(M)$, then the equations above simplify to:

$$P(l_T|\boldsymbol{X}_{1:t}) = \frac{\exp(Z_l(t))}{\sum_{l=0}^{M} \exp(Z_l(t))}.$$

The posterior on l_T is sufficient to compute likelihood of $\boldsymbol{X}_{t+1}$:

$$P(\boldsymbol{X}_{t+1}|\boldsymbol{X}_{1:t}) = P(\boldsymbol{X}_{t+1}, C=0|\boldsymbol{X}_{1:t}) + P(\boldsymbol{X}_{t+1}, C=1|\boldsymbol{X}_{1:t}),$$

$$\text{where } P(\boldsymbol{X}_{t+1}, C=0|\boldsymbol{X}_{1:t}) = P(\boldsymbol{X}_{t+1}|C=0)P(C=0|\boldsymbol{X}_{1:t}) = P(l_T=0|\boldsymbol{X}_{1:t})\prod_l P(\boldsymbol{X}^l_{t+1}|C^l=0),$$

$$\begin{aligned} P(\boldsymbol{X}_{t+1}, C=1|\boldsymbol{X}_{1:t}) &= \sum_{l_T} P(\boldsymbol{X}_{t+1}|l_T)P(l_T|\boldsymbol{X}_{1:t}) \\ &= \sum_{l_T} P(\boldsymbol{X}^{l_T}_{t+1}|C^{l_T}=1)\prod_{l\neq l_T} P(\boldsymbol{X}^l_{t+1}|C^l=0)P(l_T|\boldsymbol{X}_{1:t}). \end{aligned}$$

Translating optimal thresholds for discrimination to asymptotic thresholds for search

We discuss how to design thresholds for visual search that asymptotically achieve the best ER vs RT trade-off (as in **Conj. 1** and **Eq. 6.3** and **Eq. 6.4**). This is done by relating the asymptotically optimal visual search thresholds $\{\tau^{vs}_-, \tau^{vs}_+\}$ to two other pairs of thresholds:

- $\{\tau_-, \tau_+\}$: the optimal thresholds for discrimination with an *even* prior ratio (i.e. $P(C=1)/P(C=0)=1$),
- $\{\tau'_-, \tau'_+\}$: the optimal thresholds for discrimination with a *biased* prior ratio of $1/M$.

(I) $\{\tau^{vs}_-, \tau^{vs}_+\} = \{\tau'_-, \tau'_+\}$: the asymptotic search thresholds are identical to the discrimination threshold with a $1/M$ prior ratio. The asymptotic case is where the locations $l \neq l^*$ are absolutely sure that they do not contain any target, i.e. $Z_l(t) \to -\infty, \forall l \neq l^*$. Asymptotically (i.e. after collecting a significant amount of information) this always happens when the target is absent, and happens with probability $1/M$ when the target is present (when l^* is the target location). Therefore, the asymptotic search problem can be reduced to a visual discrimination problem with a prior ratio of $1/M$.

(II) $\{\tau'_-, \tau'_+\} + \log(1/M) = \{\tau_-, \tau_+\}$: log prior ratio causes an additive change to the optimal discrimination thresholds. Let γ_+ and $-\gamma_-$ (note that $\gamma_+, \gamma_- > 0$) be the upper and lower thresholds for visual discrimination with a prior of p for target-present. Let RT_C and ER_C be the expected response time and error rate when the stimulus type is $C \in \{0, 1\}$. The error rates, assuming the two thresholds are far

apart, are given by (see summary in [6]):

$$\begin{aligned}
\mathrm{RT}_1(\gamma_+, \gamma_-) &\approx \mathrm{RT}_1(\gamma_+) = \frac{k}{\eta}\gamma_+, \\
\mathrm{RT}_0(\gamma_+, \gamma_-) &\approx \mathrm{RT}_0(\gamma_-) = \frac{k}{\eta}\gamma_-, \\
\mathrm{ER}_1(\gamma_+, \gamma_-) &\approx \mathrm{ER}_1(\gamma_-) = \frac{1}{1+e^{\gamma_-}}, \\
\mathrm{ER}_0(\gamma_+, \gamma_-) &\approx \mathrm{ER}_0(\gamma_+) = \frac{1}{1+e^{\gamma_+}},
\end{aligned}$$

where k is an unknown constant that is inversely proportional to the drift-rate. The total risk $\mathcal{R}(\gamma_+, \gamma_-)$ is given by:

$$\mathcal{R}(\gamma_+, \gamma_-) = p\mathrm{RT}_1(\gamma_+) + (1-p)\mathrm{RT}_0(\gamma_-) + p\mathrm{ER}_1(\gamma_-) + (1-p)\mathrm{ER}_0(\gamma_+).$$

At the optimal thresholds γ_+^* and γ_-^*, it must be that the local derivatives of the risk function w.r.t. the thresholds are zero:

$$\begin{aligned}
\frac{\partial \mathcal{R}}{\partial \gamma_+}\Big|_{\gamma_+=\gamma_+^*} = 0 \implies \frac{k}{\eta} &= \frac{(1-p)e^{-\gamma_+^*}}{p(1+e^{-\gamma_+^*})^2} \approx \frac{1-p}{p}e^{-\gamma_+^*} = e^{-\left(\gamma_+^* + \log\frac{p}{1-p}\right)} \\
&\implies \gamma_+^*(p) = -\log(\frac{k}{\eta}) - \log\frac{p}{1-p}, \\
\frac{\partial \mathcal{R}}{\partial \gamma_-}\Big|_{\gamma_-=\gamma_-^*} = 0 \implies \gamma_-^*(p) &= -\log(\frac{k}{\eta}) + \log\frac{p}{1-p}.
\end{aligned}$$

Setting $p = 1/2$ (or equivalently, $\log\frac{p}{1-p} = 0$) and $p = 1/(1+M)$ (or equivalently, $\log\frac{p}{1-p} = -\log(M)$) respectively, we have:

$$\begin{aligned}
\tau_+ &= \gamma_+^*(\frac{1}{2}) = -\log(\frac{k}{\eta}), \\
\tau_+' &= \gamma_+^*(\frac{1}{1+M}) = -\log(\frac{k}{\eta}) + \log(M) \\
\implies \tau_+' &= \tau_+ + \log(M).
\end{aligned}$$

Similarly,

$$\implies \tau_-' = -\gamma_-^*(\frac{1}{1+M}) = -(\gamma_-^*(\frac{1}{2}) - \log(M)) = \tau_- + \log(M).$$

Therefore, the optimal thresholds $\{\tau_-', \tau_+'\}$ with a biased prior ratio may be obtained by offsetting the optimal thresholds $\{\tau_-, \tau_+\}$ with the log prior ratio.

Combining (I) and (II), see see that the asymptotic visual search thresholds are given by $\{\tau_-^{vs}, \tau_+^{vs}\} = \{\tau_-, \tau_+\} + \log(M)$.

References

[1] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," 2015.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (ICML)*, vol. 32, 2015, pp. 448–456.

[4] B. Chen, V. Navalpakkam, and P. Perona, "Predicting response time and error rates in visual search," in *Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2699–2707.

[5] W. J. Ma, V. Navalpakkam, J. M. Beck, R. Van Den Berg, and A. Pouget, "Behavior and neural basis of near-optimal visual search," *Nature Neuroscience*, vol. 14, no. 6, pp. 783–790, 2011.

[6] J. Palmer, A. C. Huk, and M. N. Shadlen, "The effect of stimulus strength on the speed and accuracy of a perceptual decision," *Journal of Vision*, vol. 5, no. 5, pp. 376–404, 2005.

[1]