Quantum of Vision

Thesis by
Bo Chen

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

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2016
Defended May 12, 2016

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).

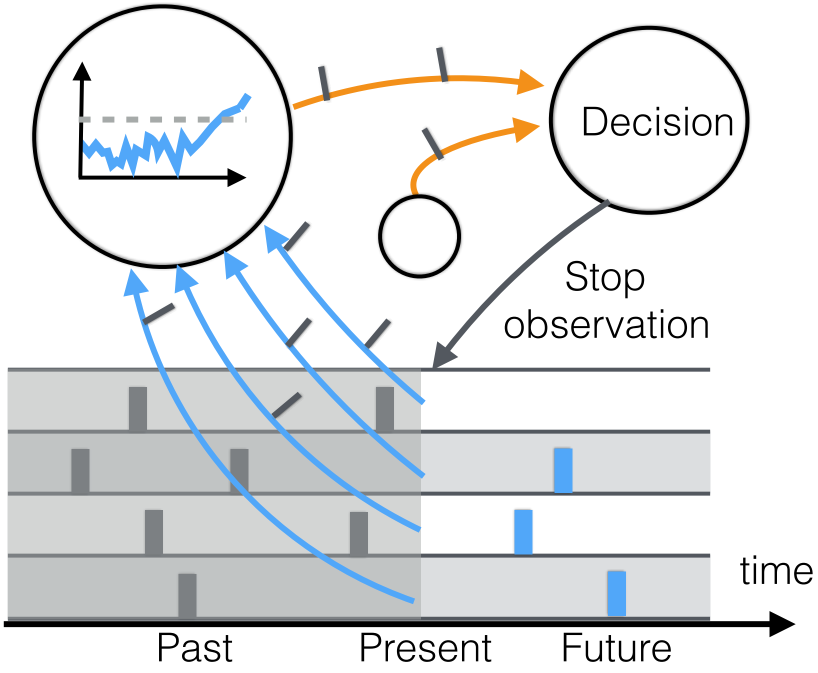

Figure 1.1: **Quantized vision**. Information trickles into a vision system (through blue arrows) one quantum at a time. The vision system may also be quantized in that computation flows through the system in small packets (through orange arrows). The quantization in the input provides flexibility to 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 $\exp(x) \triangleq 10^x$.

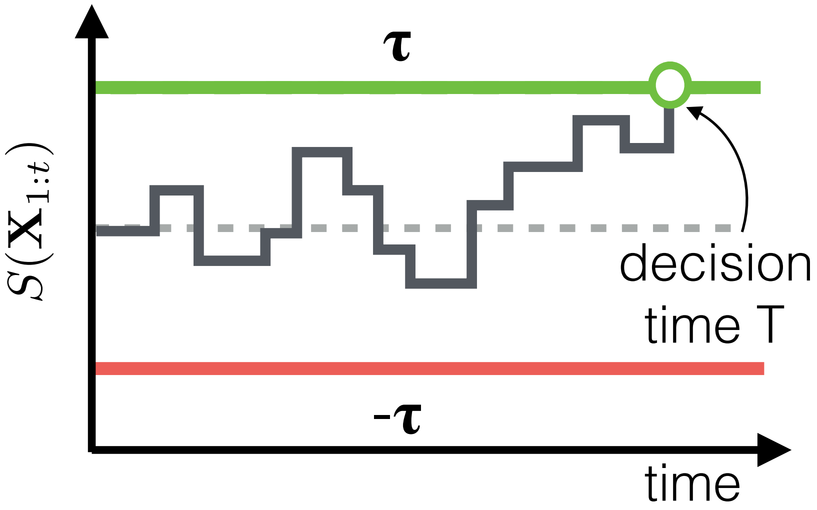

Figure 2.1: **The sequential probability ratio test (SPRT)**. SPRT **Eq. 2.3** computes the log class posterior ratio $S(X_{1:t}) = \log \frac{P(C=1|X_{1:t})}{P(C=0|X_{1:t})}$ and compares to a pair of constant thresholds (assumed symmetrical here) for deciding whether to continue collecting observations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

> Here we assume that the two classes share the same prior probability of 0.5, hence the log posterior ratio $\log P(C = 1|X_{1:t})/P(C = 0|X_{1:t})$ is identical to the log likelihood ratio $\log \frac{P(X_{1:t}|C=1)}{P(X_{1:t}|C=0)}$. 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 $S(X_{1:t})$ 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 $S(X_{1:t})$ 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.

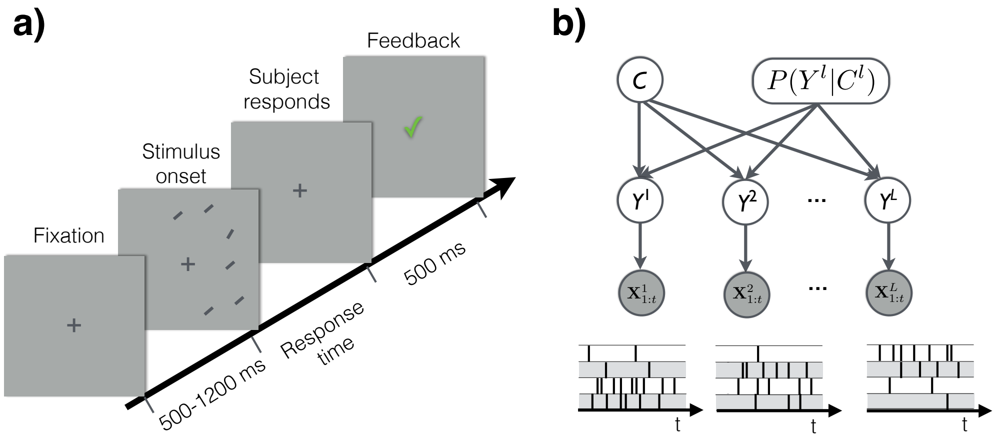

Figure 3.1: **Visual search setup** (**a**) Each trial starts with a fixation screen. Next, the “stimulus” is displayed. The stimulus is an image containing M oriented bars that are positioned in M out of L possible display locations ($M = 6$, $L = 12$ in this example). One of the bars may be the target. 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

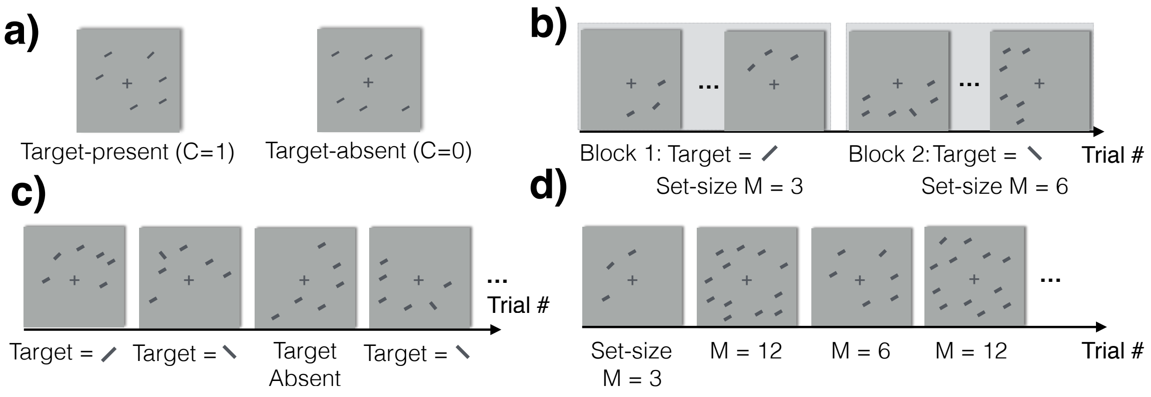

Figure 3.2: **Common visual search settings.** (**a**) The two stimulus categories: target-present and target-absent. (**b**) “Blocked”: the orientation contrast and the set-size remain constant within a block of trials (outlined in gray boxes) and vary only between blocks. (**c**) “Mix contrast”: the 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.

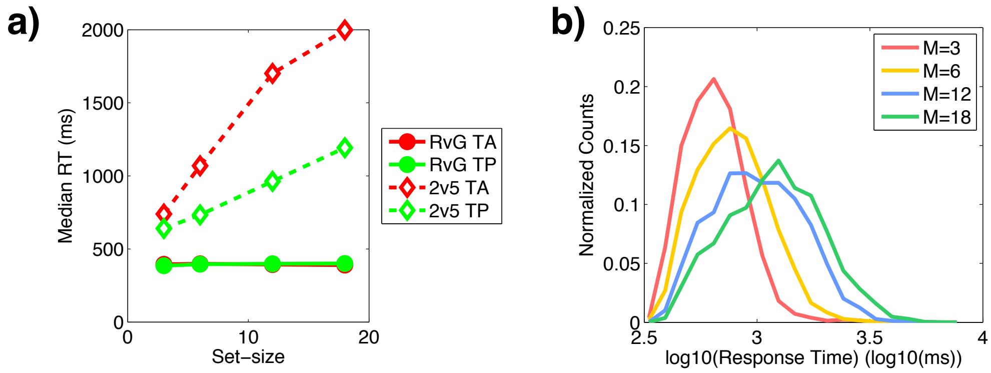

Figure 3.3: **Selected list of visual search phenomena** (**a**) The “set-size” effect. Median RT increases linearly with set-size. The slope depends on the trial type (target-absent trials have roughly twice the slope) and task difficulty. The two tasks are searching for a red bar among green bars (easy) 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_θ^i, 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.

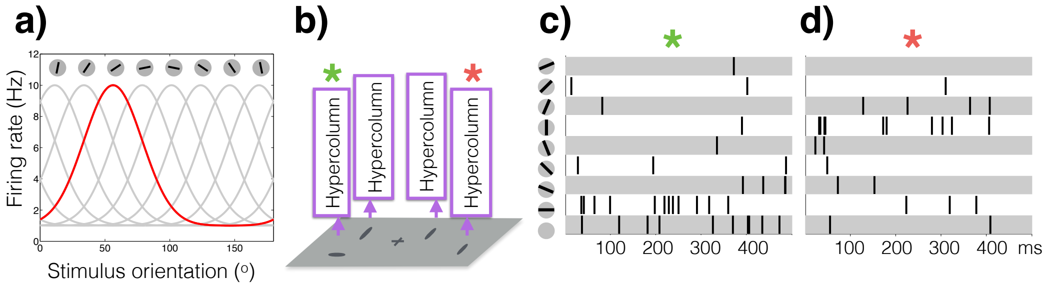

Figure 3.4: **V1 Hypercolumns** (**a**) Orientation tuning curves λ_θ^i (**Eq. 3.1**) of a hypercolumn consisting of $n_H = 8$ neurons with half tuning width $\sigma_Y = 22°$, minimum firing rate $\lambda_{min} = 1Hz$ and maximum firing rate $\lambda_{max} = 10Hz$. (**b**) V1 hypercolumns tessellate 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

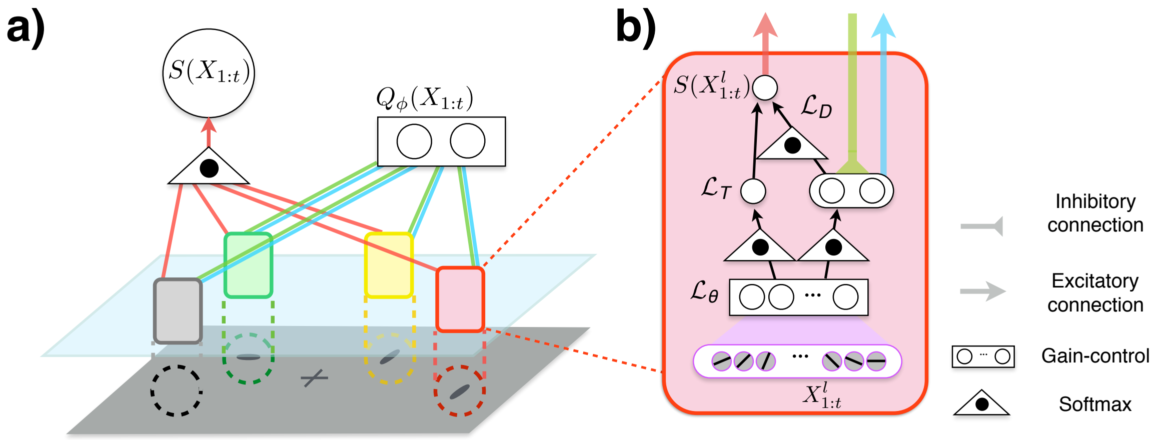

Figure 3.5: **SPRT for heterogeneous visual search.** (**a**) SPRT for heterogeneous visual search is implemented by a five-layer network. It has two global circuits, one computes the global log likelihood ratio $S(\boldsymbol{X}_{1:t})$ (**Eq. 3.10**) from local circuits that compute log likelihood 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:

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

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

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.

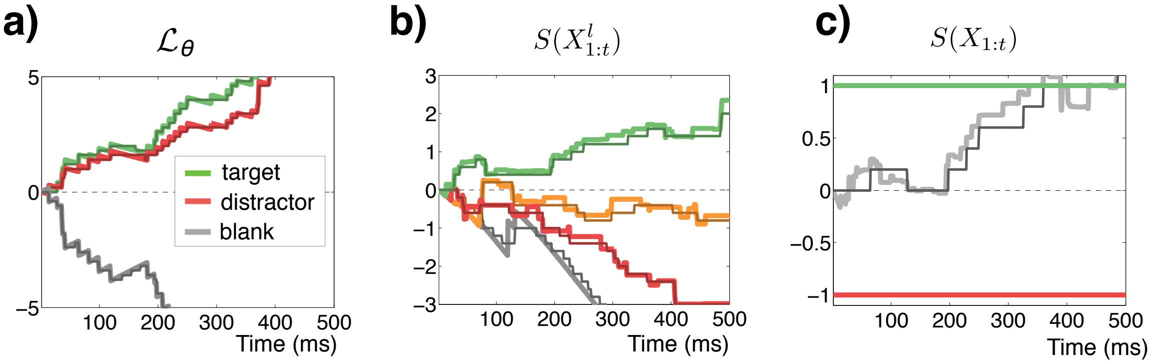

Figure 3.6: **An instantiation of the signals propagating through the network in Fig. 3.5a.** The orientation contrast is 45° and there are two possible set-sizes, 1 and 3. (**a**) The orientation log likelihoods $\mathcal{L}_\theta(\boldsymbol{X}_{1:t})$ (**Eq.** 3.13) at the target location (green box in **Fig.** 3.5**a**). 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.** 3.5**a**. (**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**):

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

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

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**).

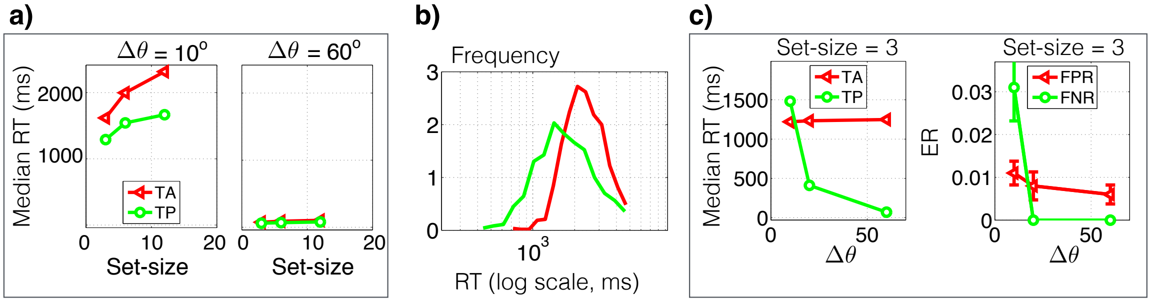

Figure 3.7: **Qualitative predictions of SPRT (Sim. 1-2).** (**a**) Set-size effect on median RT under the blocked design (**Sim. 1**). The ideal observer predicts a linear RT increase with respect to set-size when the orientation contrast $\Delta\theta$ is low (10°, left) and a constant RT when the 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-

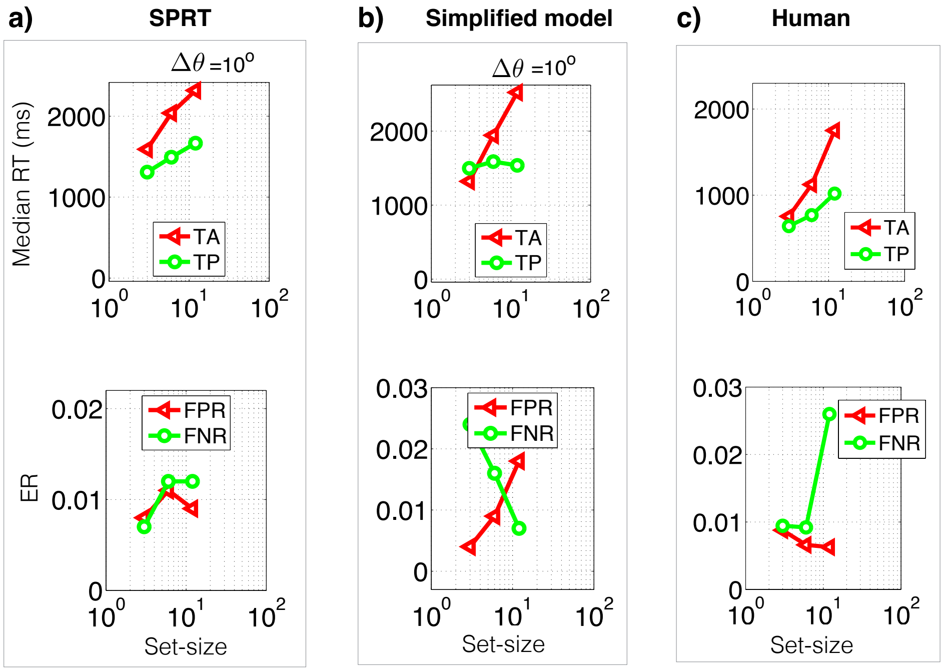

Figure 3.8: **Qualitative model predictions and psychophysics data on visual search with unknown set-size (Sim. 3).** Median RT (upper) and ER (lower): false-positive-rate (FPR) and false-negative-rate (FNR), of visual search with homogenous target/distractor and unknown set-sizes (**Sim. 3**) under 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 × 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

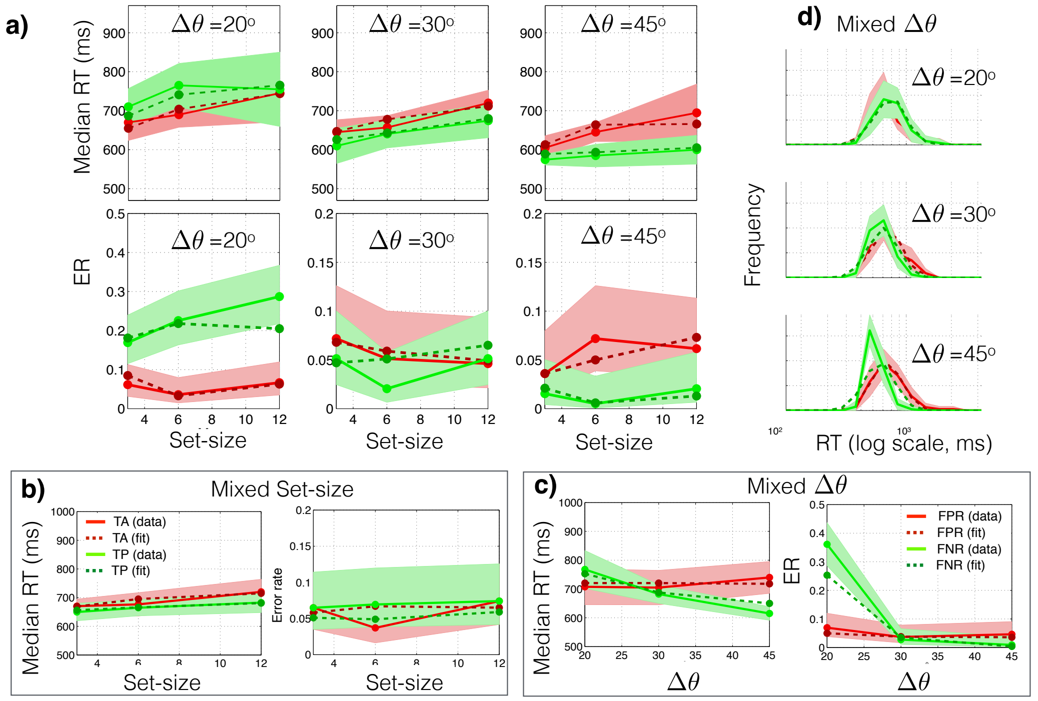

Figure 3.9: **Behavioral data of a randomly selected human subject and fits (ER, median RT and RT distributions) using SPRT**. (**a**) **Exp. 1**: “Blocked" design. All set-size M and orientation contrast $\Delta\theta$ combinations share the same hypercolumn and non-perceptual parameters; the decision 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.

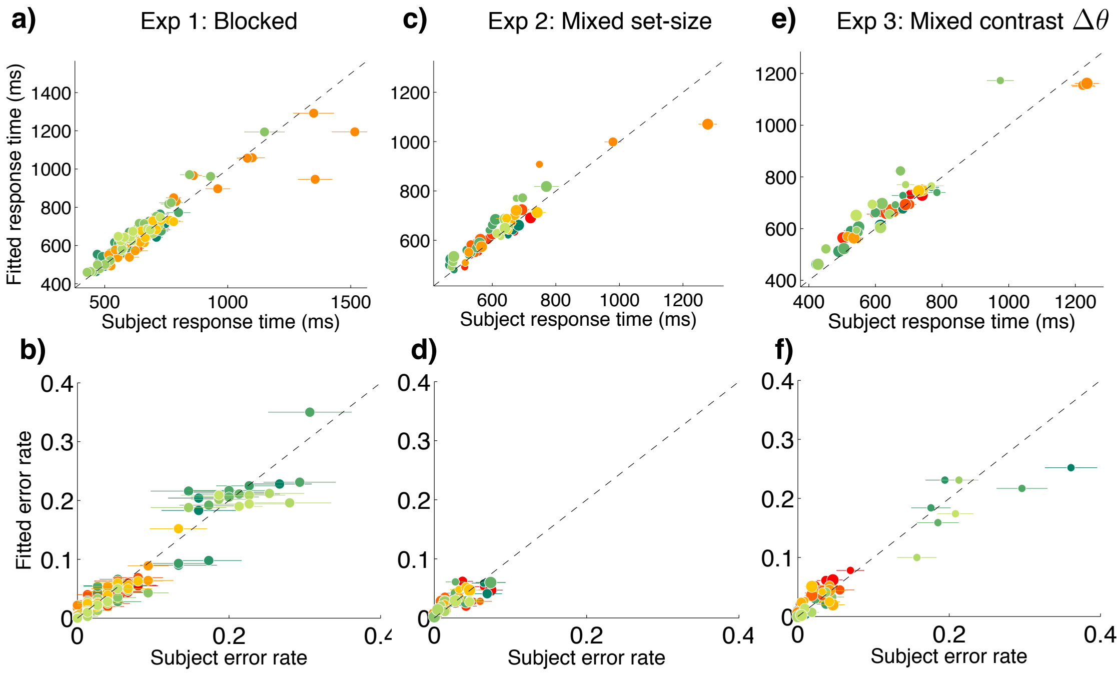

Figure 3.10: **Synopsis of fits to nine individual subjects.** The rows correspond respectively to three designs: **Exp. 1** (blocked), **Exp. 2** (mixed contrast) and **Exp. 3** (mixed set size). The maximum firing rate of the hypercolumn λ_{max} and the two non-decision parameters for each subject 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

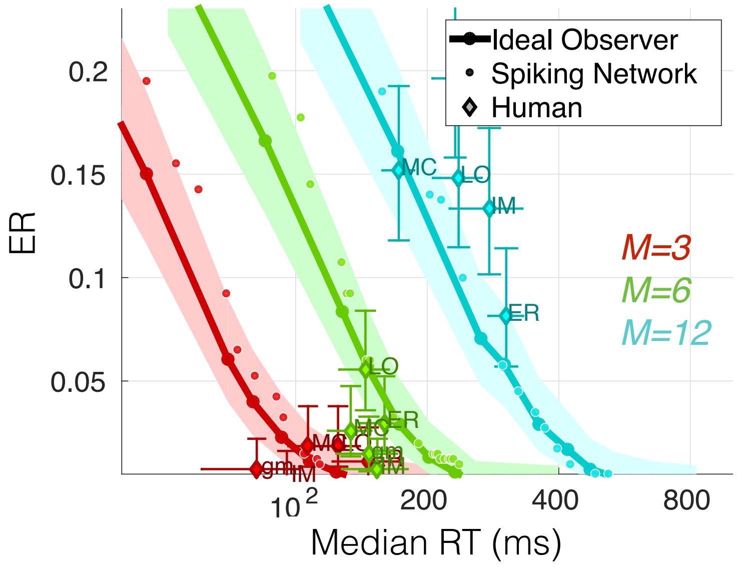

Figure 3.11: **Speed accuracy tradeoff.** ER vs RT tradeoff in the mixed set-size task (**Exp. 3**, **Fig. 3.2d**) of five human subjects (ER, IM gm, MC, and LO) with similar estimated internal parameters, as well as of SPRT (**Eq. 3.2**) and the spiking network implementation (**Sec. 3.6**) using the 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**:

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

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

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

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].

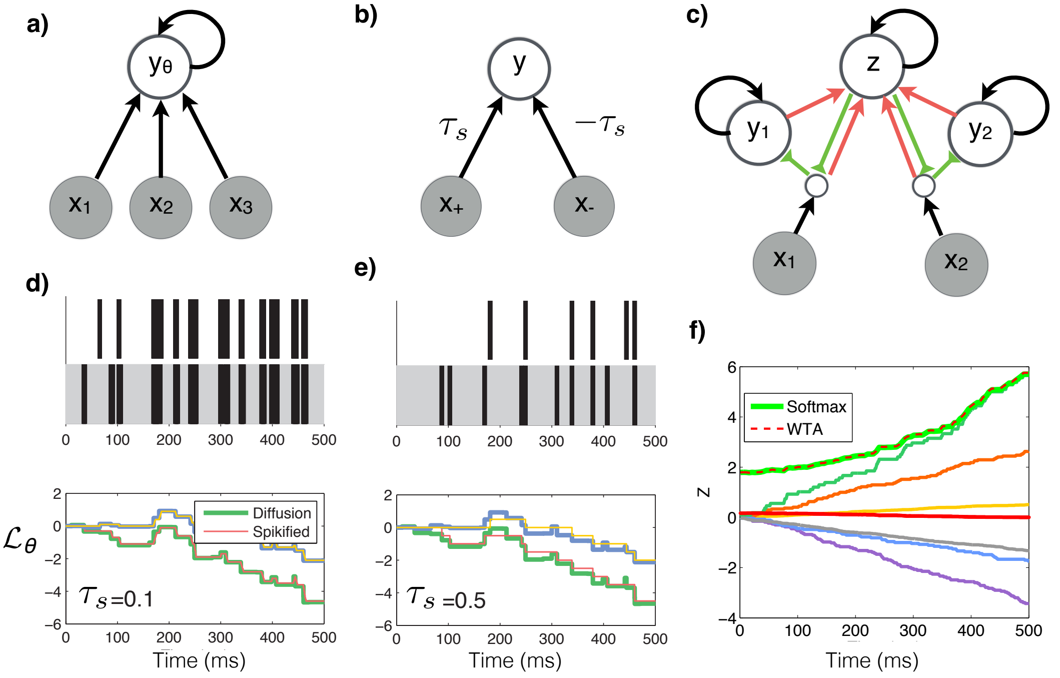

Figure 3.12: **Spiking implementation** (**a**) A feedforward network implemented the average gain-controlled network of **Eq. 3.14**. (**b**) Signal transduction. The positive and negative parts of the signal in x are encoded with integrate-and-fire neurons and transmitted to the receiver neuron y. (**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.