Chapter 6

OPTIMALITY ANALYSIS OF SEQUENTIAL PROBABILITY RATIO TEST

Strictly Optimal Sequential Tests

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

6.1 Optimal decision strategy for homogeneous search

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

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

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

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

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

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

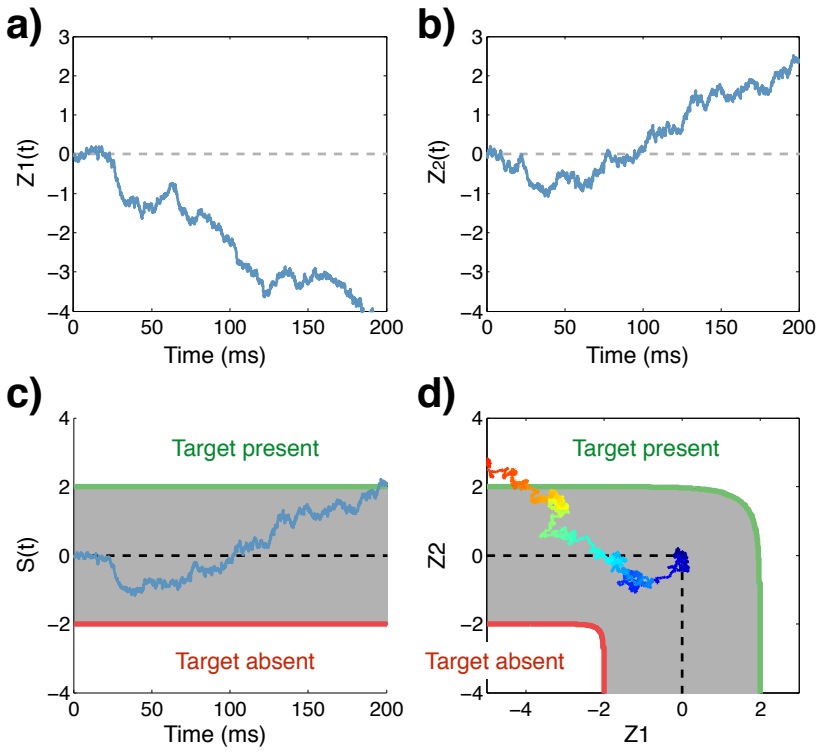

Figure 6.1: **Decision strategies for homogeneous visual search.** To perform probabilistic inference, a sequential test computes for each location the local log likelihood ratio $Z_l = \log \frac{P(X^l|C_l=1)}{P(X^l|C_l=0)}$ over time: **(a)** Z_l at a distractor location, **(b)** Z_l at the target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

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

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

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

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

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

Computational solution for low-dimensional problems

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

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

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

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

Conjecture for high-dimensional problems

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

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

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

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

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

where a_+ and a_- are unknown parameters.

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

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

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

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

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

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

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

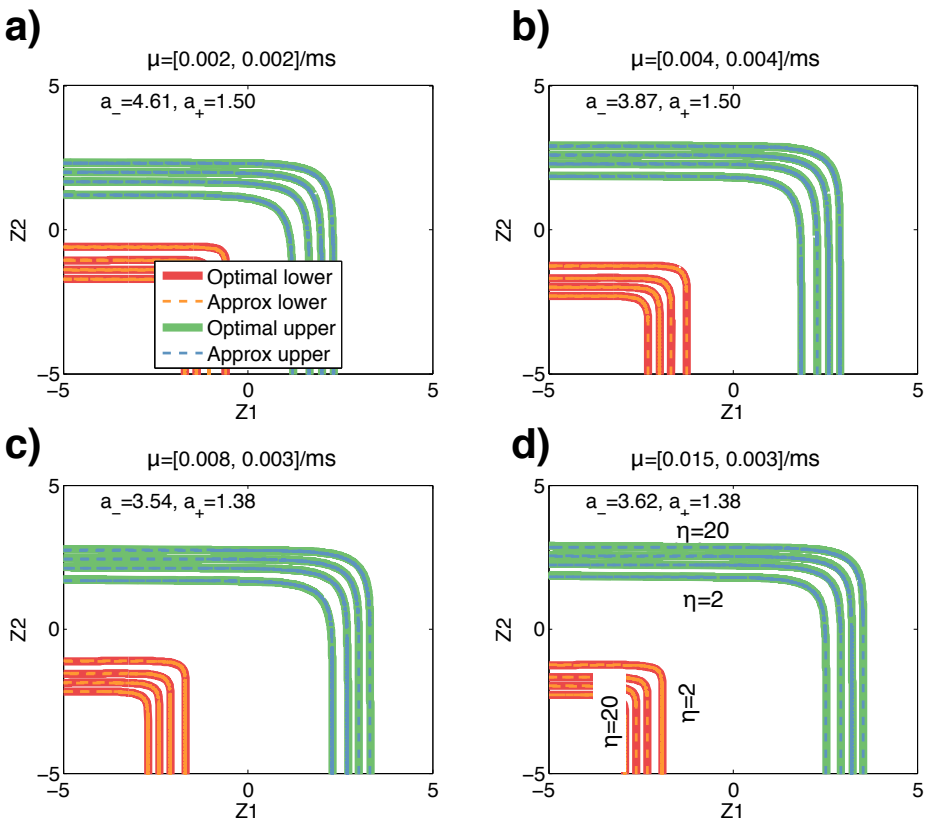

Figure 6.2: **Optimal sequential test for 2D visual search.** **(a-b)** Optimal decision thresholds and approximations for different costs of errors $\eta =\in \{2, 5, 10, 20\}$) in homogeneous search. Decision boundaries are approximated using **Eq. 6.5** and 6.6 with $a_+ = 1.50$ and $a_- = 4.61$.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

for visual search with the same time cost is:

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

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

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

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

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

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

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

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

6.2 Optimality analysis of current search models

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

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

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

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

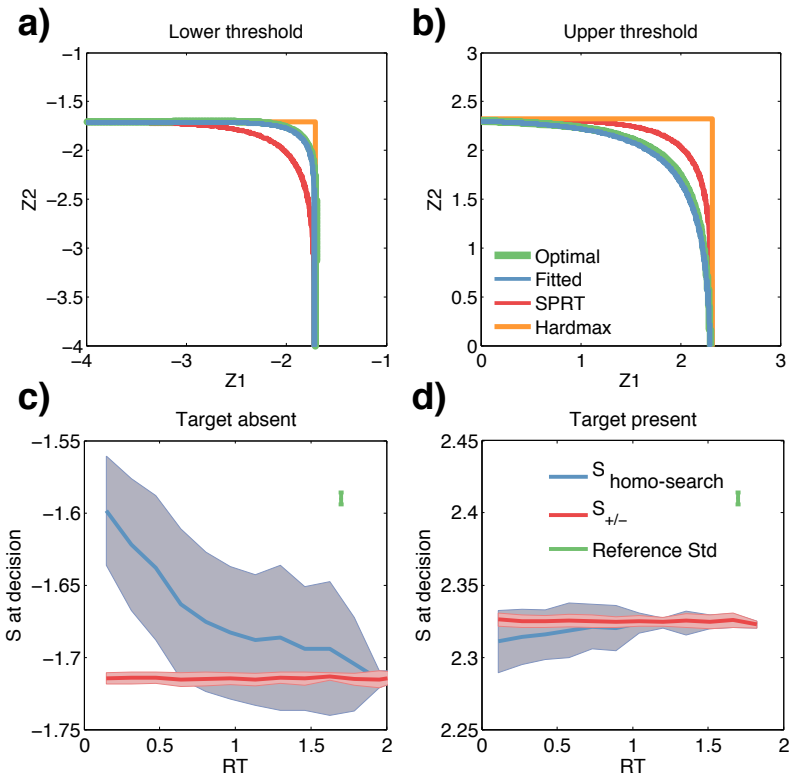

Figure 6.3: **Sequential testing strategies for homogeneous visual search in two-dimensions.** The optimal and various alternative decision strategies are compared in terms of **(a)** the lower and **(b)** the upper threshold in the joint space of $\{Z_1, Z_2\}$. The a-SPRT thresholds are obtained from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

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

Decision surfaces comparison.

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

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

Can time-varying threshold make SPRT optimal?

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

Risk comparison.

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

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

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

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

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

6.3 Chapter summary

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

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

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

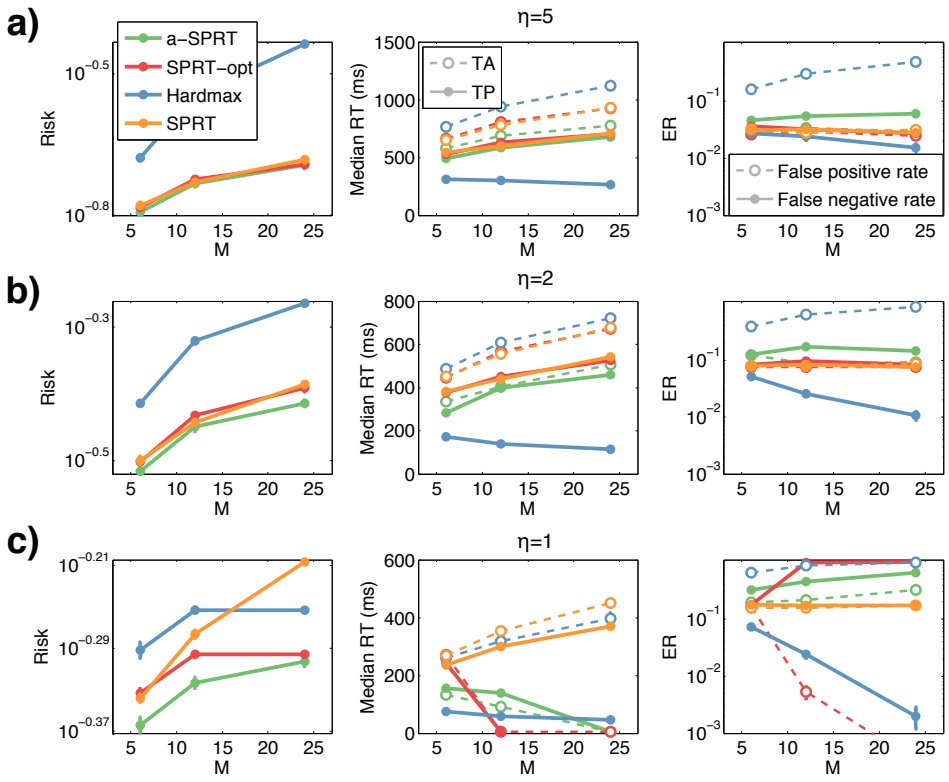

Figure 6.4: **Risk comparison of common decision strategies in homogeneous visual search.** a-SPRT, SPRT-opt, SPRT and Hardmax are compared under different costs of errors: **(a)** $\eta = 5$, **(b)** $\eta = 2$, and **(c)** $\eta = 1$ with a drift-rate of $\pm 12/sec$. Hardmax is sub-optimal in all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

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

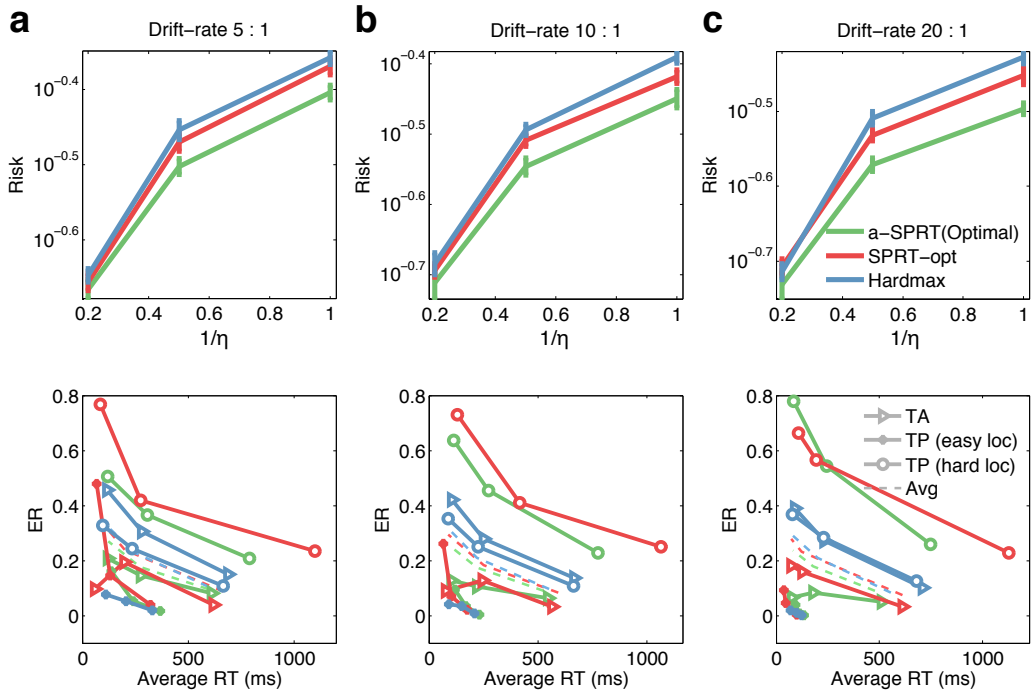

Figure 6.5: **Risk comparison of common decision strategies in heterogeneous drift-rate visual search.** a-SPRT, SPRT-opt and Hardmax are compared under various costs of time. The first row shows the overall risk versus the cost of error. The second row shows the ER vs RT tradeoff under different costs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

References

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

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

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

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

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

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

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

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

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

Chapter 7

DISCUSSION AND CONCLUSIONS

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

1

Appendix A

APPENDIX

A.1 Visual search

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

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

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

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

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

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

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

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

Review: Bayesian inference for discrimination and homogeneous search

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

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

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

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

which proves **Eq. 3.3**.

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

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

which proves **Eq. 3.5**.

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

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

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

Formulating common search problems using the general model

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

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

In this case a CDD is a three dimensional vector of

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

We will employ three CDDs:

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

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

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

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

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

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

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

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

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

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

where ϵ is a small number to prevent zero probability.

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

Bayesian inference for heterogeneous visual search

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

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

where

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

Similarly, the target-absent likelihood is:

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

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

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

Then the log likelihood ratio is

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

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

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

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

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

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

Mean-field approximation to SPRT

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

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

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

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

Search with correlated target and distractor orientations

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

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

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

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

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

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

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

A.2 Scotopic visual recognition

Time-adaptation of hidden features (Eq. 4.8)

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

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

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

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

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

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

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

which has a posterior mean of:

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

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

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

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

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

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

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

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

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

thus the log posterior ratio is:

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

which proves **Eq. 4.8**.

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

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

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

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

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

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

Datasets

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

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

Implementation

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

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

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

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

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

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

A.3 Visual discrimination with unknown stimulus onset

Log posterior ratios based on momentary observations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

When $k = 0$, we have:

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

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

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

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

Log posterior ratios based on spike trains

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

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

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

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

Similarly,

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

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

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

which admits the following recursive computation:

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

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

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

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

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

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

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

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

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

which proves **Eq. 5.13**.

A.4 Optimality analysis

State formulation in visual search

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

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

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

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

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

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

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

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

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

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

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

Translating optimal thresholds for discrimination to asymptotic thresholds for search

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

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

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

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

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

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

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

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

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

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

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

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

Similarly,

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

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

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

References

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

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

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

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

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

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

[1]