Exploring fixed-threshold and optimal policies in multi-alternative decision making

Michael Shvartsman
Princeton Neuroscience Institute
Princeton University
ms44@princeton.edu

Vaibhav Srivastava
Electrical and Computer Engineering
Michigan State University
vaibhav@egr.msu.edu

Jonathan D. Cohen
Dept. of Psychology and Princeton Neuroscience Institute
Princeton University
jdc@princeton.edu

Abstract

The dynamics of human and animal behavior within a perceptual decision made based on a single stationary stimulus are consistent with sequential statistical testing (e.g. Bogacz et al. 2006) instantiated as the discrete-time sequential probability ratio test (SPRT; Wald & Wolfowitz, 1948) or its continuous time analogue, the diffusion model (DDM; Ratcliff, 1978). In this simple domain, the SPRT/DDM with a fixed threshold is both reward-rate- and Bayes-optimal.

However, in nonstationary or multihypothesis settings, these criteria need not be equivalent: fixed threshold policies are not optimal under either criterion, and there is no systematic framework to compute reward-rate-optimal policies (though cf. Mahadevan, 1996; Dayanik & Yu, 2013). Consequently, work on the dynamics of decisions over nonstationary stimuli or multiple choices has either explored Bayes-optimal policies by dynamic programming (e.g. Frazier & Yu, 2008; Drugowitsch et al. 2012) or used fixed threshold policies (e.g. McMillen & Holmes 2006, Norris 2009).

We use our model of the dynamics of multi-stimulus decision making to explore the differences between fixed-threshold and Bayes-optimal policies in different tasks, exploiting the connections between Markov decision processes, Bayesian inference, and diffusion (e.g. Dayan & Daw, 2008) to do so. We show that even in simple tasks, predictions can depend on whether we assume the organism uses the fixed-threshold policy or the Bayes-optimal policy. Specifically, we show that different explanations for the flanker effect (Yu, et al. 2009; White et al. 2011) are normative under different choices of the action set and policy space. We additionally show that the Bayes-optimal policy makes the unusual prediction that as the posterior probability of some hypotheses drops due to evidence, the decision criterion for the remaining hypotheses should rise. We speculate that the intention superiority effect in prospective memory could be evidence of such a rise.

Keywords: MDP, POMDP, SPRT, MSPRT, optimality, rationality, dynamical systems, DDM, flanker, AX-CPT, prospective memory, threshold policy

1 Introduction

In very simple settings, it is possible to investigate the question of how closely human or animal behavior matches normatively optimal performance. In these cases – most notably, a perceptual two-alternative forced-choice (2AFC) decision based on a single stationary stimulus, human and animal behavior is close to statistically optimal under multiple different criteria [9, 10]. Beyond these simple cases, there are different ways in which constraints and goals can interact, making it difficult to distinguish between an agent performing suboptimally and one performing near-optimally but with respect to different constraints or goals. Thus, normative approaches to complex behavior tend to search for the conditions humans or animals optimize for rather than signatures for optimal behavior itself, including agent goals [1], bounds on computation [12], heuristics [7], or restrictions on the optimization algorithm itself [13].

We focus on an extension of the 2AFC setting to choices that admit multiple stimuli from perception and memory. Even this simple extension is nontrivial and an area of current research [e.g. 4, 14, 17]. Our framework, while similar to these other approaches, stands apart in (a) naturally capturing a variety of tasks of interest in cognitive psychology and neuroscience within a single framework [16] and (b) allowing multiple stimuli to be mapped to the same response, thereby decoupling the decision from stimulus identification. Even in this simple setting, the problem specification is challenging. For example: is the policy space that subjects explore unconstrained, or constrained to the space of fixed-threshold policies? Is the action space restricted to sampling all stimuli, or to sampling each stimulus separately?

To investigate this, we exploit the connections between dynamical systems, Bayesian inference, and Markov decision processes [e.g. 2] to flexibly move among equivalent formulations of the problem. We begin with Bayesian inference, use its continuum limit approximate a belief transition density, and use this transition density in a Markov decision process to derive Bayes-optimal policies. We provide a number of insights: first, we show how different proposed explanations for the classic Flanker effect can be thought of as normative consequences of different restrictions on the policy space and the action set. Next, we compute the optimal policies for the AX Continuous performance test (AX-CPT) and prospective memory (PM) tasks, two tasks in which there is not a one-to-one mapping from stimuli to responses. We show how the optimal policies are not equivalent to fixed-threshold policies on the posterior probability of the response, and identify a signature property of optimal policies for these tasks: the increase of the response criterion on one stimulus as the other stimulus or stimuli become better-identified. We speculate about empirical findings that may be a consequence of this property, including the intention superiority effect in the prospective memory literature.

2 Theoretical framework and background

Suppose that an agent is interested in inferring the true value of a noisily observed *target* stimulus G that can take on two values g_0, g_1 based on noisy i.i.d. samples s_G. It can then perform the following sequence of Bayes updates:

$$P_\tau(G = g_i) \propto P(s_G \mid G = g_i)P_{\tau-1}(G = g_i);\ i \in \{0, 1\};\ \tau > 0; \tag{1}$$

Where $P_\tau(\cdot)$ is the posterior probability of the argument being true given observations until time τ and we omit the priors $P_0(\cdot)$ for brevity. In this setting, the agent can select a threshold and stop sampling at the first point at which the posterior of either stimulus exceeds this threshold. This policy is a notational variant of the Sequential Probability Ratio Test (SPRT; [18]). For a stationary distribution of s_G, an appropriately selected threshold will simultaneously optimize at least two reward criteria: (a) reward rate, and (b) Bayes risk. If the distribution of s_G is Bernoulli, this problem is isomorphic to the Tiger problem in the study of partially-observable Markov decision processes (POMDPs) [11]. However, consider the slightly more complicated setting:

$$P_\tau(C = c_i, G = g_j) \propto P(s_C, s_G \mid C = c_i, G = g_j)P_{\tau-1}(C = c_i, G = g_j); \quad i \in \{0, 1\}, j \in \{0, 1\} \tag{2}$$

We have added a second stimulus C which we deem the *context* to distinguish it from the target stimulus G. This context stimulus can be perceptual, or a previously encoded stimulus being retrieved from memory. In the latter case, we assume the context stimulus is retrieved from memory with decay modeled by an AR(1) process (see [16] for detail).

Within this framework, a task is defined by two properties: first, the onset and offset times of the stimuli; second, a mapping from the joint identity of the two stimuli to the correct response. For example, if the stimuli appear simultaneously and the task goal is to identify the target stimulus G while ignoring the context, the task is the Flanker or Stroop. If the stimuli appear asynchronously and the goal is to make one response to c_0 trials regardless of G and otherwise identify the G stimulus, then the task is the prospective memory task [5]. Other tasks can be represented similarly, and in this way our model is an abstract model of multi-stimulus decision making applicable to a variety of tasks.

To identify the true context and target pair, we could define a fixed threshold over their joint density, which would make this model equivalent to the multihypothesis sequential probability ratio test [MSPRT; 3]. The MSPRT is an asymptotically optimal test in the sense that it approximates the Bayes-optimal test as accuracy approaches 100%. However, such rule may lose optimality guarantees if multiple pairs of context and stimuli map to the same decision. Alternatively, a fixed-threshold decision rule can be designed by setting a fixed threshold on the response probability computed by appropriately combining context-target probabilities using the true response rule.

This is a sensible heuristic because the space of fixed threshold policies over the response posterior (henceforth: FTR policies) is low-dimensional and therefore relatively easy to explore quickly, and was used in past work to simulate plausible behavior patterns from multiple tasks under the same parameterization [16]. In contrast, the Bayes-risk optimal (henceforth: BRO) policies must be defined over the full belief simplex [3]. Presently a fully analytic way leading to a closed-form optimal test is not available, and in fact it may not, in general, exist. However, the question within neuroscience is far from settled, with some recent work suggesting that subjects do use or approximate nonstationary policies, specifically ones that either accelerate the sampling rate or lower the threshold as time passes [4, 17].

We explore the consequence of the FTR and BRO policies for a number of different tasks implemented using the framework described above. To do so, we re-specify the problem as a discrete-time POMDP for which we can find optimal policies by standard methods. The states are the four possible identities of the stimulus pair ($[c_i, g_j], i \in \{0,1\}, j \in \{0,1\}$) and the actions are to sample both stimuli, respond left, and respond right. In some cases, we also add the actions to sample s_C and s_G independently. The transitions are trivial in that the sampling action stays at the current stimulus state and the response actions end the episode. The rewards are specified in terms of costs of -1 for a sampling either stimulus a -40 for an error response. Sampling both stimuli simultaneously costs -1.5, since we assume cost includes opportunity cost and effort cost, and the former is not duplicated when sampling both together. The specific parameter values are not important for our claims – outside of degenerate cases, the qualitative patterns are similar over the space.

We intentionally leave the observation density unspecified for a reason that will become clear shortly. First, we redefine the problem as a belief-MDP parameterized by the pair of log-likelihood ratios z_c, z_g. In the continuum limit, this model is equivalent to the following two-dimensional Ornstein-Uhlenbeck process over this belief space [16]:

$$\mathrm{d}\vec{z} = (\vec{a} - \Lambda\vec{z})\mathrm{d}t + \Sigma \mathrm{d}\vec{W}, \quad \vec{a} = \begin{pmatrix} a_c \\ a_g \end{pmatrix}, \quad \vec{\Lambda} = \begin{pmatrix} \lambda_c & \lambda_{cg} \\ \lambda_{gc} & \lambda_g \end{pmatrix} \quad z_c := \log \frac{P_\tau(C = c_0)}{P_\tau(C = c_1)} \quad z_g := \log \frac{P_\tau(G = g_0)}{P_\tau(G = g_1)} \tag{3}$$

The drift vector $\vec{a}$ contains the expected value of log-likelihood ratios of the evidence for both stimuli, and Σ is the covariance matrix of those ratios. $\mathrm{d}\vec{W}$ are i.i.d. increments from a Wiener process, the diagonal entries of Λ are AR coefficients, and its off-diagonal elements are mutual excitation terms that will be non-zero if $P(s_C, s_G \mid C, G) \neq P(s_C \mid C)P(s_G \mid G)$. We can set Σ to be the identity matrix w.l.o.g. because it is indeterminate with Λ. This formulation is advantageous because (a) we only specify $\vec{a}$ and Λ rather than observations and their probabilities, and (b) transition densities for the sampling actions are mixtures of Gaussians with means $\pm\vec{a} - \Lambda\vec{z}$ and identity covariance. We discretize time at $\mathrm{d}t = 1$. We discretize the belief space to a 110×110 grid between ± 3 in both dimensions, and solve it by backward induction in discrete time with the horizon set well beyond typical decision times.

3 The Flanker task

The Flanker task [6] requires subjects to identify a central target stimulus (e.g. $>$ or $<$) in the presence of interfering flankers that may be either congruent or incongruent with the target (e.g. $>>>>>$ vs. $<<><<$). Subjects are slower and less accurate on incongruent trials, and guess below chance on fast responses [8]. This effect has been explained in sequential sampling models by making one of three different assumptions: (a) that perceptual uncertainty makes samples from target and flankers correlated [20]; (b) that subjects have a bias to expect congruent trials [20]; and (c) that a shrinking attentional spotlight admits flanker samples into decisions early but not late [19]. We show how all three solutions can be optimal under different conditions. We model the central target stimulus as G and the surrounding flankers as C, with the reward function reflecting the true response rule. Specific parameters values are not critical to our claims – here we used 0.3 and 0.1 for context and target drifts (since there are more flankers than target).

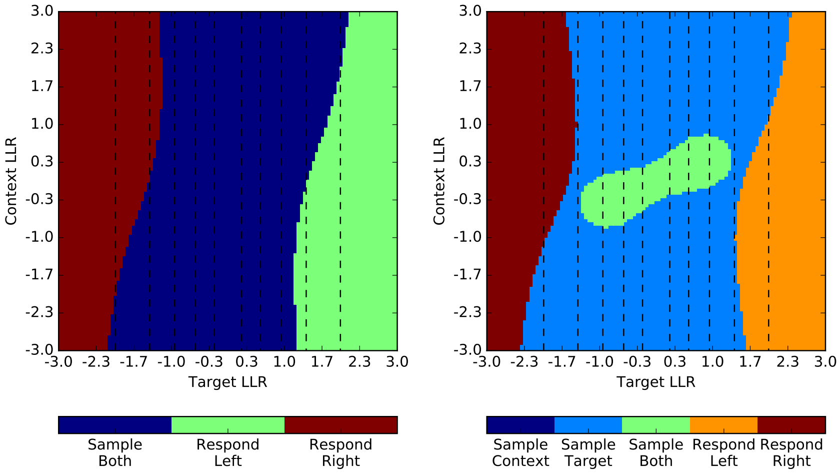

Figure 1: **Optimal policies for flanker task**. *Left*: Optimal policy for flanker with spatial uncertainty and obligatory sampling of both stimuli (model 1). *Right*: same model with ability to sample stimuli independently (model 3). Dashed contours mark possible FTR policies.

The first model we consider has no ability to direct its attention among the stimuli, but implements spatial uncertainty by setting the off-diagonal terms of Λ to -0.3. As Fig. 1 (left) shows, the optimal policy for this model is not an FTR policy. Since congruent samples move the belief faster towards decision, the value of sampling is higher in the congruent quadrants, which moves the decision boundary farther in the congruent parts of the belief space. This compensation will reduce or eliminate the effect of congruence on behavior relative to the FTR policy, which cannot make this adjustment to the decision boundary. This makes the spatial uncertainty explanation more compatible with fixed-threshold than with BRO policies. The second model we consider (not shown) sets the spatial uncertainty to zero, which makes the BRO and FTR policies equivalent. This model is only compatible with the congruence-bias explanation, under which the agent is likelier to start in the top-right and bottom-left quadrants, and therefore be faster to respond and more accurate on congruent trials (this explanation is compatible with all the models we explored).

The third model we consider (Fig 1, right) combines spatial uncertainty with the ability to selectively sample from either of the stimuli. This model chooses the sample-both action when uncertainty is highest (center of the space), and switches to sampling only the target once the flankers have been somewhat identified. However, the switch happens at higher LLR in congruent than incongruent trials, because flanker samples are more useful in congruent trials. In this way, the model has a rudimentary adaptively shrinking attentional spotlight [19]. In sum, all three previous claims about explanations for flanker effects can be framed as normative claims, under different policy spaces and action sets.

4 The AX-CPT and Prospective Memory tasks

Next we tackle two tasks in which the response rules are more complicated than simply marginalizing over one stimulus. The first is the symmetric AX Continuous Performance Test (AX-CPT) [15]. In the symmetric AX-CPT, participants see one of two context stimuli (by convention labeled 'A' or 'B') followed by one of two targets ('X' or 'Y'), and make one response (e.g. 'left') to AX and BY pairs, and the other (e.g. 'right') to AY and BX pairs. This task is among the simplest that require the composition of stimuli to make a decision, and as such is a useful test-bed for understanding more complex decision making. To model this task in our framework we treat the first stimulus ('A' or 'B') as context and the other as target, and use the true response rule for the reward function. We set both drifts to 0.2 and the context decay portion of Λ to 0.1 (as before, the results are not strongly dependent on these parameters).

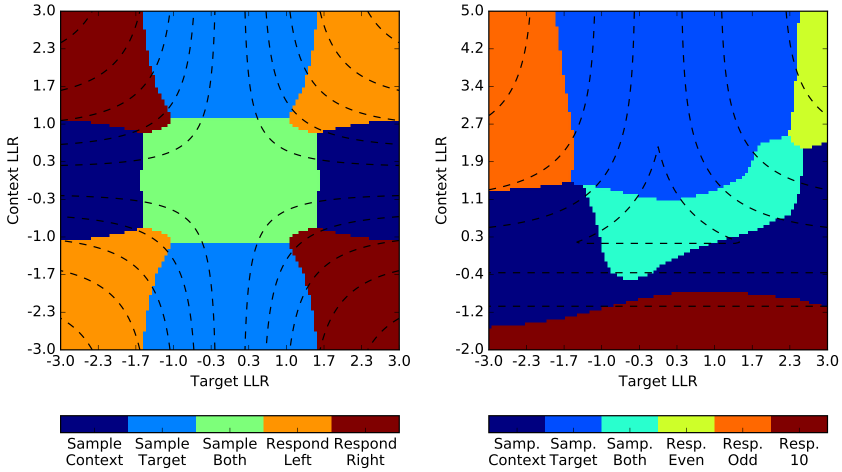

Figure 2: **Optimal policy for AX-CPT (left) and Prospective Memory (right).** Dashed contours mark possible FTR policies.

Fig. 2 (left) shows the optimal policy for this task. The BRO agent samples both stimuli until it has enough evidence for one, and then samples just the other. With respect to the decision, the BRO policy is once more not contained in the space of FTR policies. Rather, it makes the prediction that the decision bounds w.r.t. one stimulus widen as the posterior of the other concentrates. As above, this is again because the value of sampling increases – this time due to the structure of the task rather than interference. The intuition is that if the agent is sure that the first stimulus is 'A', then, distinguishing between 'X' and 'Y' is valuable, since that will allow a correct response. If the agent is uncertain about the context stimulus, distinguishing between the target stimuli is less useful, because responses based on one stimulus alone will have high error probability (up to 50% if all trial types are equally likely). This threshold increase also appears in the BRO (but not FTR) policy for the general multihypothesis setting [3]. As such, it is a candidate prediction for distinguishing between BRO and FTR policies. We suspect that a stimulus noise manipulation in AX-CPT could be used to test this prediction: more noisy signal of one stimulus should result in faster decisions. This is counterintuitive because typically one would expect a more difficult stimulus to result in a slowdown rather than a speedup.

The final task we investigate is the prospective memory task [5], which, we speculate, may provide a evidence for this signature of BRO policies. This task investigates the ability to maintain long-term goals in the face of an ongoing task by asking subjects to make a sequence of simple two-alternative choices (e.g. 'is a number odd or even?') while keeping in mind an additional task goal (e.g. 'press a third button if the number is 10'). The *intention superiority* effect in this task is the fact that correct task responses are faster when the prospective cue is present. That is, subjects are faster to respond that 10 is even than that other numbers are even.

We model the task as a three-alternative forced choice, with the context samples reflecting the long-term task ('is the number 10?') and the target samples reflecting the ongoing task. The parameters are identical to those used in the AX-CPT model, except for the addition of the uncertainty between the context and target (because 10 is also even), and a larger grid in the context dimension to better illustrate the boundary widening. As Fig. 2 (right) shows, the BRO policy in this task has narrower decision bounds on the parity task if the 10-detection evidence is stronger. This means, counterintuitively, that the threshold on declaring a stimulus as even is lower when the stimulus is '10', which may provide a normative explanation for the intention superiority speedup noted above.

5 Discussion and conclusion

We have explored he difference between assuming a fixed response threshold and computing fully Bayes-optimal policies for a number of different tasks that can be expressed within our framework for modeling multi-stimulus decisions. We also explored the impact of the available action sets on predicted behavior. We have shown that the differences between Bayes-optimal and fixed-threshold policies yield qualitative differences in predictions across a number of tasks, and speculated on settings where these predictions may be tested. Future work will investigate these predictions, and explore the subtleties of optimal policies in explaining human behavior. On the methodological front, we have demonstrated the power of exploiting equivalences between the probabilistic, MDP, and dynamical systems formulations of the same problem: equivalences that are not necessarily new, but not frequently used in either the RL or mathematical psychology communities. We believe that this approach can be further exploited in the future to continue bridging these related approaches to understanding decision-making.

References

[1] Anderson, J. R. (1990). *The adaptive character of thought*. Routledge.

[2] Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

[3] Dragalin, V., Tartakovsky, A. G., and Veeravalli, V. (1999). Multihypothesis sequential probability ratio tests .I. Asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461.

[4] Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, 32(11):3612–28.

[5] Einstein, G. O. and McDaniel, M. a. (2005). Prospective Memory. Multiple Retrieval Processes. *Current Directions in Psychological Science*, 14(6):286–290.

[6] Eriksen, B. A. and Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1):143–149.

[7] Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

[8] Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., and Donchin, E. (1988). Pre- and poststimulus activation of response channels: a psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):331–344.

[9] Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psycho-physics*. Wiley, New York.

[10] Holmes, P. and Cohen, J. D. (2014). Optimality and some of its discontents: successes and shortcomings of existing models for binary decisions. *Topics in cognitive science*, 6(2):258–78.

[11] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134.

[12] Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, 6(2):279–311.

[13] Lieder, F., Hamrick, J. B., Hay, N. J., Plunkett, D., Russell, S. J., and Griffiths, T. L. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. *Advances in Neural Information Processing Systems*, pages 2870–2878.

[14] Norris, D. (2009). Putting it all together: a unified account of word recognition and reaction-time distributions. *Psychological Review*, 116(1):207–19.

[15] Servan-Schreiber, D., Cohen, J. D., and Steingard, S. (1996). Schizophrenic Deficits in the Processing of Context. *Archives of General Psychiatry*, 53(12):1105.

[16] Shvartsman, M., Srivastava, V., and Cohen, J. D. (2015). A Theory of Decision Making Under Dynamic Context. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2485–2493. Curran Associates, Inc.

[17] Thura, D., Beauregard-Racine, J., Fradet, C.-W., and Cisek, P. (2012). Decision-making by urgency-gating: theory and experimental support. *Journal of Neurophysiology*.

[18] Wald, A. and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326–339.

[19] White, C. N., Ratcliff, R., and Starns, J. J. (2011). Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cognitive psychology*, 63(4):210–38.

[20] Yu, A. J., Dayan, P., and Cohen, J. D. (2009). Dynamics of attentional selection under conflict: toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, 35(3):700–17.