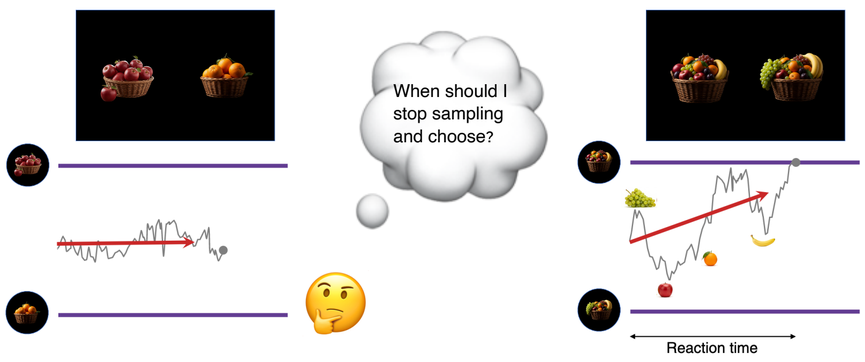
RNNs Reveal a New Optimal Stopping Rule in Sequential Sampling for Decision-Making

Jialin Li, Kenway Louie, Paul Glimcher, Bo Shen

NYU Grossman School of Medicine, New York, NY

Motivation

- In naturalistic environments, evidence streams are usually **heterogeneous**
- Agent might be resource-rational when sampling incurs **cost**



Homogeneous Stream **Heterogeneous Stream**

In this work, we trained RNNs to investigate stopping strategies when sampling heterogeneous evidence.

Sequential Probability Ratio Test

Wald (1945) formulated sequential sampling as a hypothesis testing problem, where the objective function is to minimize the samples taken when subject to a certain false rate α and β.

$$\min_{S \in \mathcal{S}} \quad \pi_0 \mathbb{E}_0[n] + \pi_1 \mathbb{E}_1[n]$$
$$\text{s.t.} \quad \Pr_0^S(\text{refuse } H_0) \le \alpha, \qquad \Pr_1^S(\text{accept } H_0) \le \beta$$

The decision threshold can be written as

$$a \approx \log \frac{\beta}{1-\alpha}, b \approx \log \frac{\alpha}{1-\beta}$$

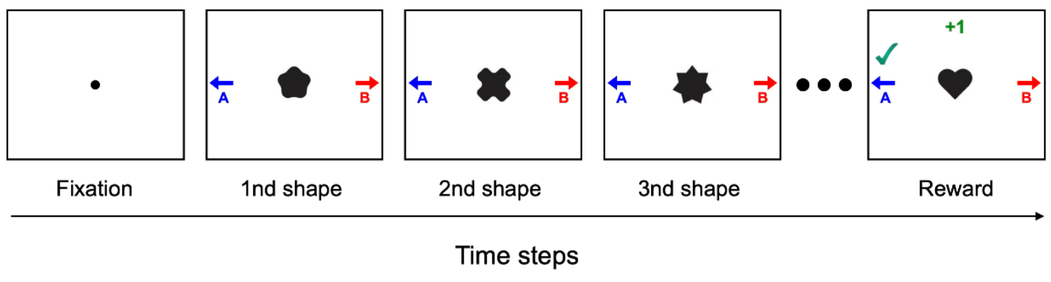
At each time step, the agent accumulate evidence based on the log-likelihood as ratio

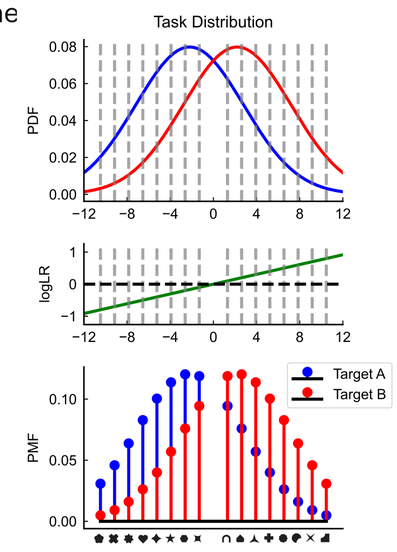
$$\ell_n = \sum_{t=1}^{n} \ln \frac{f_1(x_t)}{f_0(x_t)}$$

The decision is made until ℓ_n reaches the decision threshold

Sequential Sampling Inference Task

- Hidden target (A or B) randomly selected per trial, determining the sampling frequency of the stimuli.
- Agent receives one stimulus at a time sequentially from the given Gaussian distribution (*heterogeneous evidence*).
- Task for the agent: deciding to choose A/B or continue sampling (cost c) at each time



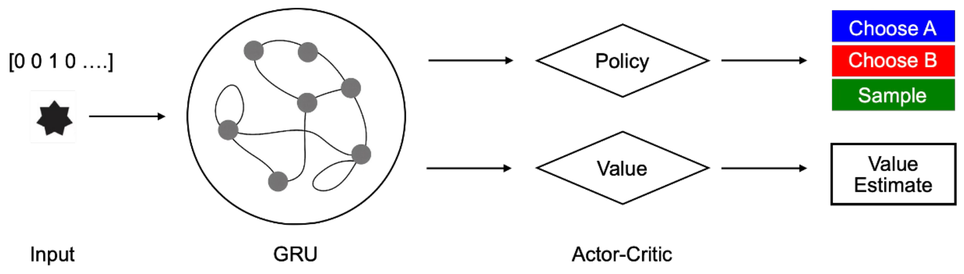


To study the stopping rule in heterogeneous environment, we tested two environmental factors:

- Sampling cost (5 levels): 0.01, 0.02, 0.03, 0.04, 0.05
- Time constraints (2 levels): No time constraint / time constraint at time step 10

Model Architecture

The model consists of GRU cell with 64 hidden units, a policy head and a value head.



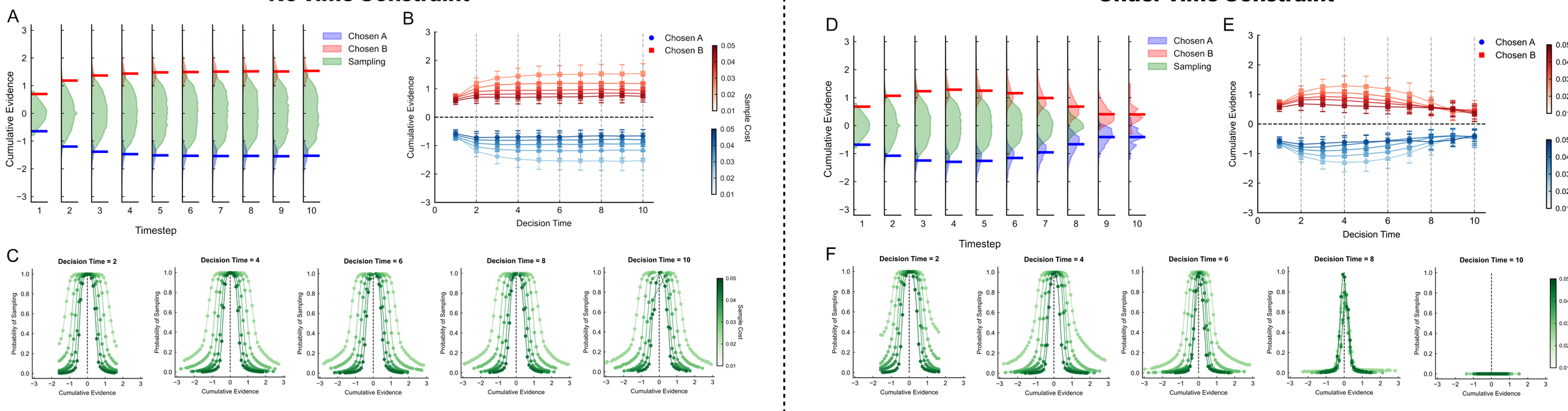
The models were trained on-policy using an **advantage actor-critic (A2C) algorithm**:

$$J(\theta, \phi) = \mathbb{E}_t \left[\log \pi_\theta(\boldsymbol{a} \mid s_t) A_t + \beta_v \left(R_t - V_\phi(s_t) \right)^2 + \beta_e \mathscr{H}(\pi(\cdot \mid s_t)) \right]$$

where π_θ is the decision policy, s_t is the current state given the observations seen so far at step t, $\boldsymbol{a}$ are the possible actions, $A_t = R_t - V_\phi(s_t)$ is the advantage function between the actual reward R_t and the reward expectation $V_\phi(s_t)$.

- The policy (actor) term encourages the network to take actions to maximize returns.
- The value term (critic) trains the network to predict the amount of return.
- The entropy term encourages exploration behavior to prevent the network from being trapped in local minima.

Early Commitment and Collapsing Boundaries Emerged in Heterogeneous Evidence Environment



Normative Analysis

We formalized the sequential sampling problem under the resource-rational framework as a Meta-level Markov Decision Process (Meta-MDP) that optimizes decision quality under sampling and urgency costs.

At each step, the agent maintains a belief $b_t = P(H_1 \mid x_{1:t})$ over two hypotheses H_0, H_1. When agent receive a new sample, it update the posterior through:

$$b_t = p(H_1 \mid x_{1:t}) = \frac{p(x_t \mid H_1) \cdot b_{t-1}}{p(x_t)} = \frac{f_1(x_t) \cdot b_{t-1}}{f_1(x_t) \cdot b_{t-1} + f_0(x_t) \cdot (1 - b_{t-1})}$$

To derive the optimal policy, the bellman optimality is used to calculate the value of each action:

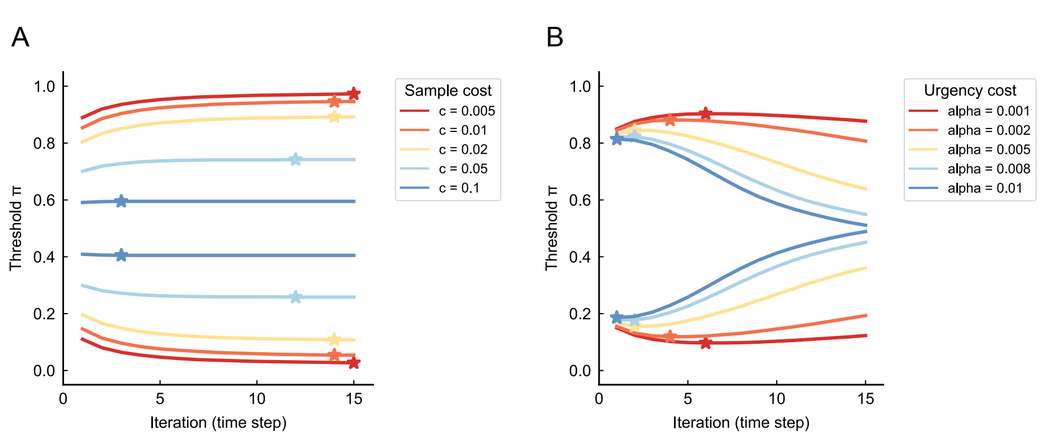
$$V^*(b,t) = \max \left\{ \underbrace{R_0(b_t, \perp)}_{\text{Choose } H_0}, \underbrace{R_1(b_t, \perp)}_{\text{Choose } H_1}, \underbrace{-c + \mathbb{E}_{x|b}\left[V(b'(x))\right]}_{\text{Sample}} \right\}$$

Where we assume that the value of sampling is given by iterating over every possible belief in the next time step t+1:

$$\mathbb{E}_{x|b}\left[V_{t+\Delta t}\left(b'_{t+\Delta t}(x)\right)\right] = \int_{-\infty}^{\infty} V_{t+\Delta t}\left(b'_{t+\Delta t}(x)\right) f(x \mid b)\, dx$$

To capture the urgency cost induced by time constraints, we introduce an extra urgency coefficient α increasing over time: $c_{total} = c_{sample} + \alpha t$

Our normative analysis reproduces the decision threshold pattern shown in environment 1 and 2.



Conclusion & Future Work

- We identified an early commitment as a novel stopping strategy other than the well-known collapsing boundary in heterogeneous environment.
- These two effects are driven by distinct mechanisms, sampling cost and time constraints
- Further analyses are needed to better interpret the connection between neural network behavior, normative analysis and human behavior.

References

Kira, S., Yang, T., & Shadlen, M. N. (2015). A neural implementation of Wald's sequential probability ratio test. *Neuron*, *85*(4), 861-873.

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature communications*, *7*(1), 12400.

Wald, A. (1945). Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, *46*(2), 265-280.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

NEURAL INFORMATION PROCESSING SYSTEMS