Rats optimize reward rate & learning speed in a 2-AFC

Javier Masís, Andrew Saxe & David D. Cox
Department of Molecular and Cellular Biology & Center for Brain Science
Harvard University

Center for Brain Science

1 Overview & Motivation

Rats resemble human performance in a two-alternative forced choice (2-AFC) task

- Here we present evidence showing that rats resemble human performance (Zacksenhouse *et al.*, 2010) in a free response two-alternative forced choice task:
  - -a subset of the animals lies on or near the optimal performance curve (OPC)
  - -most bias above the curve, rather than below it
  - -the lower their error rate, the more likely the animals are near the OPC

Rats optimize more than just reward rate during learning

- We found that during learning the rats followed a non-optimal path, signaling their policy must include more than greedy optimization of reward rate.
- Our preliminary model is that the rats optimize both reward rate and learning speed, behaving sub-optimally in order to learn faster and optimize future rewards.

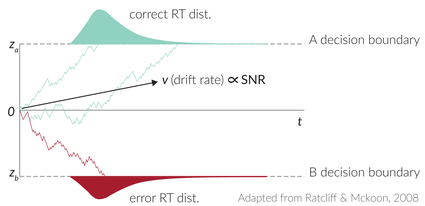
Slowing down at the beginning may make it easier to become optimal later

- Our study aims to answer the question of how agents become optimal.
- We present a theoretical model showing that going slower in early trials can increase learning speed and allow an agent to become optimal faster compared to greedily optimizing reward rate from the start.

2 Background

What is the drift difussion model?

- Drift diffusion is a model developed by Roger Ratcliff (1978) used to explain behavior in free response two-choice decision tasks. It correctly predicts accuracy and reaction time (RT) distributions in these experimental tasks in humans.
- Assumptions:
  - -noisy process that accumulates evidence over time
  - -fast decisions, made as soon as a certain threshold is crossed

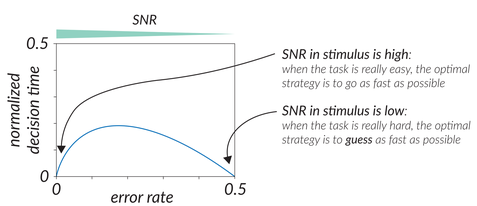


Adapted from Ratcliff & Mckoon, 2008

- How it works:
  - -An agent sees a stimulus and begins accumulating evidence about that stimulus.
  - -The rate of evidence accumulation is called the drift rate (v) (the easier the stimulus is to classify, the larger the drift rate).
  - -If the drift rate is positive, the agent will diffuse towards a positive decision boundary, and vice versa.
  - -However, there is noise involved in this process, so a positive drift rate could end up at the negative decision boundary, producing an error.

What is the optimal performance curve (OPC)?

- It is possible to determine an optimal performance curve for free response two-choice decision tasks based on the drift diffusion model (Bogacz *et al.*, 2006).
- Assumptions:
  - -agent is optimizing reward rate over anything else
  - -experimental time is fixed and trials begin immediately after the last



- How it works:
  - -The OPC predicts an optimal decision time for error rates from 0 to 0.5.
  - -This is based on the fact that there is a speed-accuracy trade-off.
  - -The decision time is scaled by the time cost of an error trial, so if getting a trial wrong means a long timeout, then the agent would benefit from going more slowly, in order to avoid making an error as much as possible.

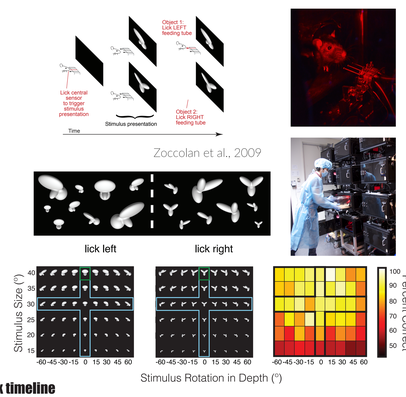
Speed-accuracy & exploration-exploitation trade-offs

- The drift diffusion model assumes evidence accumulation over time, which means there is a trade-off between how fast an agent goes and how right it is.
- There is also a trade-off between exploiting a current policy (such as guessing) and exploring new policies, which would mean forfeiting present rewards, but potentially learning policies that would yield a higher reward rate in the future.

3 Behavioral Task

Rats participate in a free response visual object recognition 2-AFC

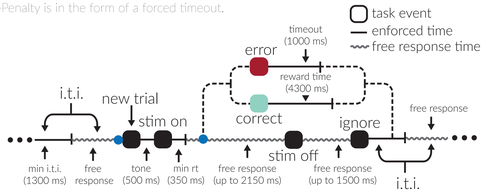
- We trained rats on an established high-throughput visual object recognition task (Zoccolan *et al.*, 2009).
- Rats are placed in behavior boxes where there are three capacitive lick ports and a screen for object presentation.
- The animals lick the center port to initiate a trial, an object is presented, and the rats choose whether to lick the right or left port depending on the object identity.
- Rats perform very well on this visual object recognition task.



Zoccolan et al., 2009

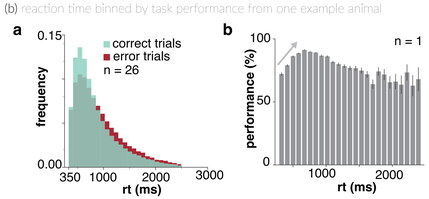
Task timeline

- Rats initiate trials at their own leisure.
- There is a minimum RT to ensure sensory processing.
- Penalty is in the form of a forced timeout.



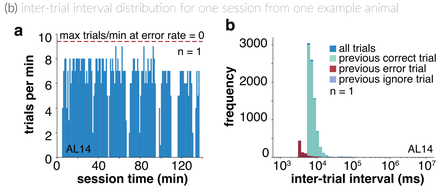
Task can be modeled using drift diffusion

- We first verified whether the task could be modeled with drift diffusion. RTs were distributed exponentially and there was a relationship between RT and performance:
  - (a) reaction time histogram for 10 latest sessions of 26 trained animals
  - (b) reaction time binned by task performance from one example animal



Free trial initiation can be approximated as forced trial initiation

- The OPC assumes trials are started one immediately after the other. Our task naturalistically allows the animals to initiate at leisure. Nonetheless, animals initiate trials at a consistent pace and work in bouts, and start most trials as soon as they can.
  - (a) histogram of trials per minute for one session from one example animal
  - (b) inter-trial interval distribution for one session from one example animal

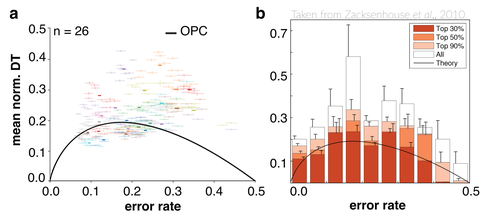


4 Results

Trained rats behave optimally & resemble human performance patterns

- Rat performance resembles human performance (Zacksenhouse *et al.*, 2010) in a free response two-alternative forced choice task:
  - -a subset lies on or near the OPC
  - -most are above the OPC
  - -the lower the error rate, the more likely they are near the OPC

- We plotted the performance of 26 trained rats spanning two training cohorts and compared their performance to human performance:
  - (a) last ten sessions of 26 trained animals in speed-accuracy space
  - (b) human performance on drifting dots experiment in speed-accuracy space



- The OPC requires normalization of decision time (DT) (Holmes & Cohen, 2014) according to the formula below, where we measured or knew ⟨RT⟩ and D_{RSI} and assumed an average T_0 (non-decision component of reaction time) of 150 ms.

$$\text{mean normalized } \langle DT \rangle = \frac{\langle DT \rangle}{D_{TOT}} = \frac{\langle RT \rangle - T_0}{T_0 + D_{RSI}}$$

⟨DT⟩ : mean DT
T_0 : non-decision time
⟨RT⟩ : mean RT
D_{RSI} : response-to-stim interval
D_{TOT} : all "other" time

Theories for suboptimal performance

- As with humans, most of our animals lay above the OPC. This is classified as sub-optimal, as subjects are taking too long to decide for a given error rate.

- There are two main hypotheses for why only a subset of humans behave optimally, and why most lie above the OPC (Bogacz *et al.*, 2006, Zacksenhouse *et al.*, 2010):
  - -there is an accuracy bias (beyond just optimizing reward rate)
  - -there is substantial timing uncertainty, and biasing to longer decision times is advantageous to maximizing reward rate when there is timing uncertainty

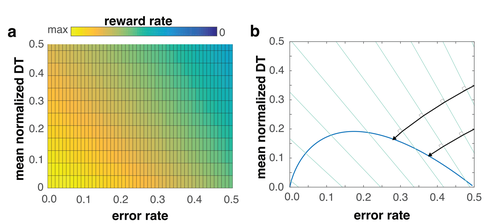
- However these theories assume a fixed performance, and do not account for learning.

- Once we saw rats could behave near optimality, we asked how it is they got there.

Null hypothesis: rats optimize reward rate during learning

- The null hypothesis in this case is to assume that the rats, like an optimal agent, are maximizing reward rate.

- To this end, we plotted iso-reward rate contours in speed-accuracy space:
  - (a) reward rate as a function of decision time and error rate
  - (b) iso-reward rate contours potted in speed-accuracy space; black arrows indicate movement maximizing the reward rate gradient



- If an agent moves perpendicular to the iso-reward rate contours (largest gradient) in speed-accuracy space, the trajectory seems to move towards the OPC. However, this would be the case anywhere in the space, such as near ER = 0.5, and mean normalized DT = 0.0. If we consider this extreme case we realize that an agent cannot move freely in this space.

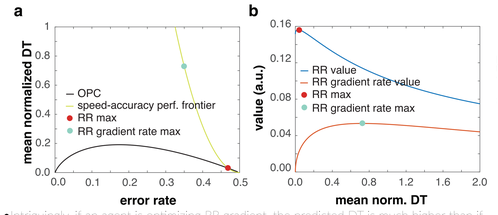
- If ER = 0.5, that means that the SNR is very low (the stimuli are indistinguishable). If the agent is responding very quickly, then there is little chance for evidence accumulation, and thus the SNR (as a stand-in for learning) should not change very much, in effect creating an area of speed-accuracy space where the agent is "stuck."

Rats may be optimizing reward rate gradient

- To formalize our intuition that an agent cannot move freely in this space, we plotted a performance frontier for a particular SNR. What the performance frontier means is that for a given SNR, an agent can only act somewhere along that curve.

- Where along that curve an agent chooses to act (effectively choosing a mean decision time for the task) will depend on the agent's learning strategy. For example, if the agent chose to maximize reward rate, then the agent would act where the performance frontier and the OPC intersect.

- Given we were curious about whether the rats may be optimizing the reward rate gradient, we calculated the value of the reward rate and the value of the reward rate gradient versus decision time:
  - (a) performance frontier in speed accuracy space
  - (b) value of RR and RR gradient according to mean normalized decision time for the performance frontier plotted in a



- Intriguingly, if an agent is optimizing RR gradient, the predicted DT is much higher than if optimizing RR. If an agent has a slower DT, this would also arguably enable faster learning.

- Thus, the prediction is that for a time early in learning, optimizing RR is at odds with optimizing learning speed. We next simulated the full dynamics to verify this prediction.

Model hypotheses for trajectory to near-optimal behavior

- We sought to understand the rats' learning trajectory to near-optimality and devised 3 models to simulate performance using a deep linear neural network.

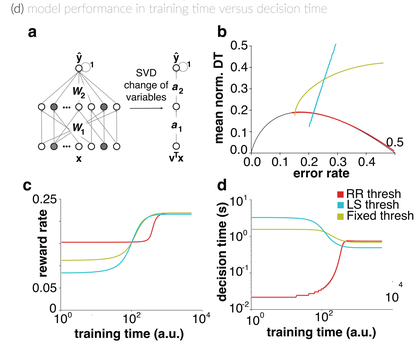
- Model description: nosiy perceptual inputs pass through two layers of tunable synaptic connections (W_1, W_2) representing the perceptual processing hierarchy, which then feeds into a perfect neural integrator representing decision-making structures. A decision is made when this activity crosses a threshold z. Because of the linearity of the system, this reduces to a drift diffusion model with an effective SNR determined by the synaptic weights. We derived a reduction of the learning dynamics in this setting when the perceptual network is trained using error-corrective gradient descent learning:

$$\tau \dot{a_i} = ER \cdot RR \cdot \frac{2\tau A (a_i c_i^2 - a_i^3 a_j^2 c_i^2)}{((a_i a_j c_i)^2 + c_o^2)^2}$$

where a_i, a_j are scalars encoding the perceptual sensitivity in each layer. ER, RR are error and reward rate, and A, c_i^2, c_o^2 are input mean, input noise variance and output noise variance respectively.

- Models:
  - RR threshold: greedily optimize instantaneous reward rate (RR)
  - LS threshold: adopt threshold that maximizes learning speed (LS)
  - Fixed threshold: adopt threshold that is optimal for performance expected after learning

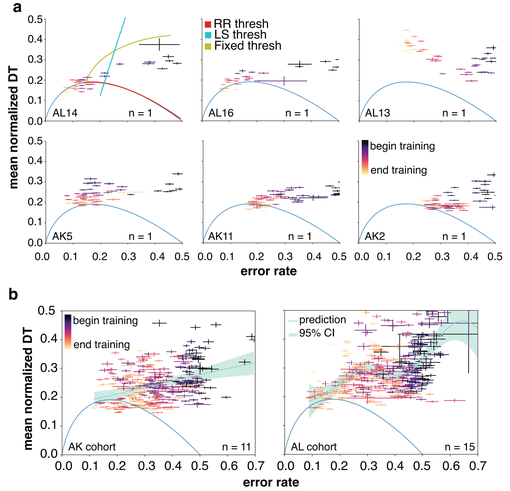
- LS & Fixed threshold yield less reward at the start but reach optimality in fewer timesteps:
  - (a) graphical representation of the deep linear neural network
  - (b) model performance in speed-accuracy space
  - (c) model performance in training time versus reward rate
  - (d) model performance in training time versus decision time



Rats appear to maximize reward rate and learning speed during learning

- We plotted the learning trajectories of 26 rats in two separate cohorts in speed-accuracy space and found that:
  - (1) rats have high RTs for high ERs, in contrast to the OPC that predicts low RTs for high ERs
  - (2) RT decreased as ERs decreased (signaling learning), although the rate of decrease in mean normalized DT was much slower than that of ER decrease.

- This trajectory best matches our Fixed threshold hypothesis, predicting that the rats are slow at the beginning with the expectation of improving future reward rate.
  - (a) learning trajectory of 6 example animals from both cohorts. Sessions are plotted from dark (early) to light (late) and have bootstrapped error bars
  - (b) average learning trajectory of both trained cohorts. We scaled session number for every animal because sessions till full training varied. We then ran a Gaussian Process to visualize the average learning trajectory for each cohort.



6 Preliminary conclusions

- Rats decidedly choose large decision times compared to the OPC when they begin learning (and error rate is near 0.5).

- Our modeling indicates there is an advantage to starting with large decision times, as learning speed is much faster, and near-optimal performance may be reached sooner.

- We are continuing to explore this phenomenon and will incorporate several experimental manipulations, such as training rats with a different error penalty, and manipulating the SNR of the stimulus in order to observe their movements in speed-accuracy space.

7 Predictions & Future Experiments

- If an agent must have fast RTs, then we predict learning speed (LS) would decrease. This can be tested by training animals with a max enforced RT.

- Model requires a prediction from the agent of how much there is to learn. Thus, if we conduct an experiment where we verbally alter expectations about how much information there is in the stimulus, we may observe predictable changes in speed-accuracy space.

- When subjects end up above OPC, are they stuck or still learning? We can alter the task parameters, such as difficulty, and penalty and observe changes in speed-accuracy space.

8 Additional Info

Acknowledgments & Funding

- Funding from the Richard A. and Susan F. Smith Family Foundation and IARPA (contract #D16PC00002).
- A.S. was supported by the Swartz Program in Theoretical Neuroscience at Harvard University. .

Email me:
jmasis@fas.harvard.edu

Find out more about the Cox Lab:
http://coxlab.org

References

- Ratcliff, R. *et al.* A theory of memory retrieval. *Psych Rev.* 85(2): 59-108, 1978.
- Bogacz, R., *et al.* The physics of optimal decision-making. *Psych Rev.* 113(4): 700-765, 2006.
- Zoccolan, D. *et al.* A rodent model for the study of invariant visual object recognition. *PNAS.* 106(21): 8748-8753, 2009.
- Zacksenhouse, M. *et al.* Robust versus optimal strategies for 2-AFC tasks. *J Math Psych.* 54(2): 230-246, 2010.
- Holmes & Cohen. Optimality and some of its discontents. *Topics Cog Sci.* 6(2): 258-278, 2014.