Dynamic Allocation Of Limited Memory Resources In Reinforcement Learning

Nisheet Patel, Luigi Acerbi, Alexandre Pouget

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- Biological agents have limited capacity to process and store information
- Standard models do not consider constraints that biological agents face:
  - Finite memory resources
  - Not all memories can be encoded with high precision

In this work, we address the following questions:

- How can agents represent values with limited memory?
- How can they utilize imprecise memories to compute good policies?
- Can they prioritize some memories over others by devoting more resources to encode them with higher precision?

Framework

We consider problems that can be characterized by a Markov Decision Process in which a resource-constrained agent represents the value of each state action pair, the q-value, with finite precision. The q-values are represented in memory as implicit distributions rather than exact estimates. This process is meant to mimic the stochastic nature of memory retrieval in noisy neural circuits.

Memories

Agents have tabular memories stored as a tuple indexed by the state and action with an imprecise estimate of the q-value: $\langle s, a, r, s', \mathcal{N}(\bar{q}_{sa}, \sigma^2_{sa})\rangle$

Policy

Agents can only draw samples from their memory distribution of q-values. If the agent is greedy, they will then choose the action corresponding to the largest sampled q-value, effectively yielding the policy:

$$\pi(a|s) = \Pr(a = \arg\max_{a'} \tilde{q}(s,a')); \qquad \tilde{q}(s,a') \sim \mathcal{N}\big(\bar{q}(s,a'), \sigma^2(s,a')\big)$$

Objective

We extend the standard RL framework such that agents maximize:

$$\mathcal{F} := \underbrace{\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+1}\middle| Q = \mathcal{N}(\bar{\boldsymbol{q}}, \boldsymbol{\sigma}^2 I)\right]}_{\text{expected future reward}} - \underbrace{\lambda D_{\mathrm{KL}}\Big(Q = \mathcal{N}(\bar{\boldsymbol{q}}, \boldsymbol{\sigma}^2 I)\big|\big|P = \mathcal{N}(\bar{\boldsymbol{q}}, \sigma^2_{\mathrm{base}} I)\Big)}_{\text{cost of representing memories precisely}}$$

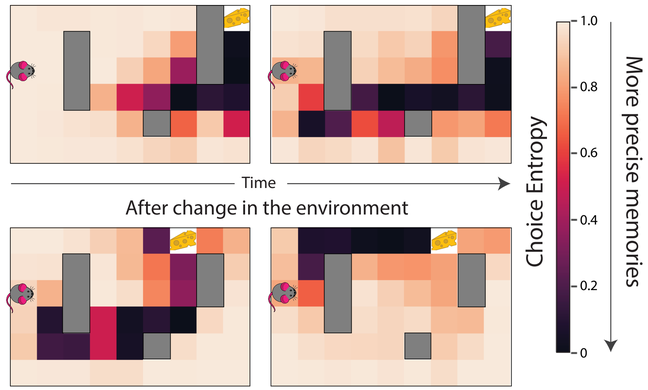
Key contributions

- We provide the agent with control over the precision of its memories, $\boldsymbol{\sigma}$.
- We derive a stochastic gradient of the objective, $\nabla_{\boldsymbol{\sigma}}\mathcal{F}$, using the policy - gradient theorem for the policy defined above.
- We combine this manner of resource allocation with learning.

Dynamic Resource Allocation

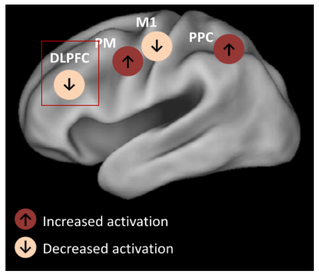
Results on a 2D gridworld

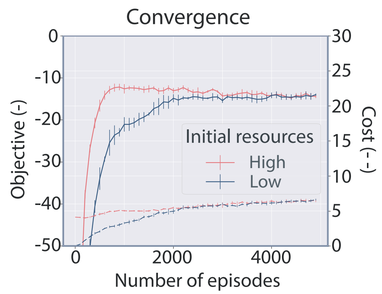
In this task[1], we can visualize the precision of the memories by looking at the entropy of the agent's choices. Lower entropy or less random choices indicate more precise memories that are shown in darker colors. Agents gradually learn to remember the shortest path to the reward even as the environment changes.



Predictions: a strategy for quicker learning

Sensory areas recruit more neurons as learning progresses, but higher-level cognitive areas (e.g. dlPFC, relevant for accessing and storing memories) lay off neurons as training progresses, implying that the task can be carried out by fewer neurons than initially recruited[2]. We show that resource-limited agents can accelerate learning by starting with high initial resources while converging to the same final solution.





References and code repository

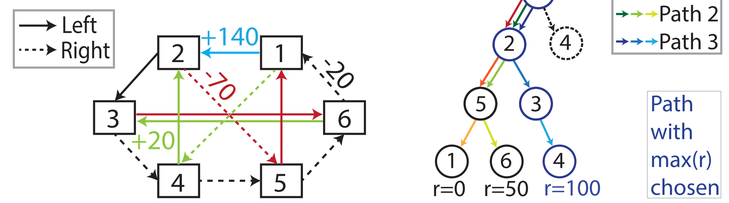
[1]Mattar and Daw (2019). Nature Neuroscience, 21(11), 1609-1617.
[2]Dayan and Cohen (2011). Neuron, 72(3), 443-454.
[3]Huys et al. (2015). PNAS, 112(10), 3098-3103.

Poster design by Jeenal Patel (jeenalpatel29@gmail.com)

Model-based Planning

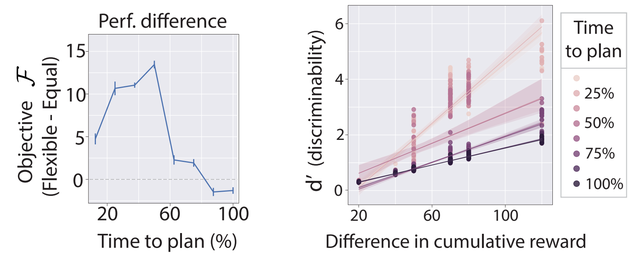
Task details

In this task[3], model-based agents are required to plan a sequence of N=3 moves that maximizes their reward in a limited time period before executing the sequence of actions. For instance, agents can draw samples of next states according to their policy until the time limit, and then choose the most rewarding planned sequence of N=3 moves.



Flexible resource allocation for better performance under time pressure

We show that it is advantageous to prioritize some memories over others by comparing our algorithm's performance against an agent constrained to have all of its memories equally precise. As expected, this advantage diminishes if agents have time to plan through all possible paths. Moreover, we show that the memories that are remembered more precisely and hence, are more discriminable (high d'), are the ones that have a larger impact on their cumulative rewards.



Conclusions

- We propose a general-purpose, dynamical framework to maximize reward under constraints of limited memory resources.
- We derive from first principles an algorithm, DRA, that learns to allocate costly resources to uncertain memories in a manner that adapts to changes in the environment.
- We provide an explanation for why frontal cortical areas in biological brains appear more engaged in early stages of learning before settling to lower asymptotic levels of activity.