Sweeping Improvements to Exploration

Georgy Antonov[1, 2, *], Peter Dayan[1, 3]

[1]Max Planck Institute for Biological Cybernetics;
[2]Graduate Training Centre for Neuroscience, University of Tubingen;
[3]University of Tubingen; *georgy.antonov@tuebingen.mpg.de

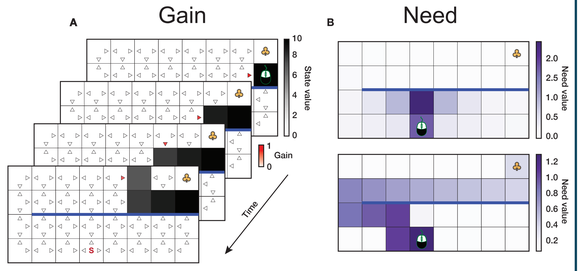
MAX PLANCK INSTITUTE FOR BIOLOGICAL CYBERNETICS

1 Background

A recent normative theory of hippocampal replay[1] suggests that the specific order of replay experiences is optimised for offline planning, whereby each replay corresponds to an update to a state-action value[2,3]

The derived prioritisation scheme considers the expected improvement in the animal's immediately ensuing behaviour as a result of an individual replay update:

$$v_{\pi_{new}}(s) - v_{\pi_{old}}(s) = \underbrace{\sum_{i=0}^{\infty} \gamma^i P(s \to x, i, \pi_{old})}_{\text{Need}} \times \underbrace{\sum_a [\pi_{new}(a \mid x) - \pi_{old}(a \mid s)] q_{\pi_{new}}(x, a)}_{\text{Gain}}$$

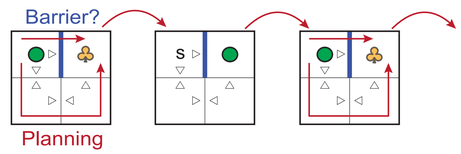


2 Problem

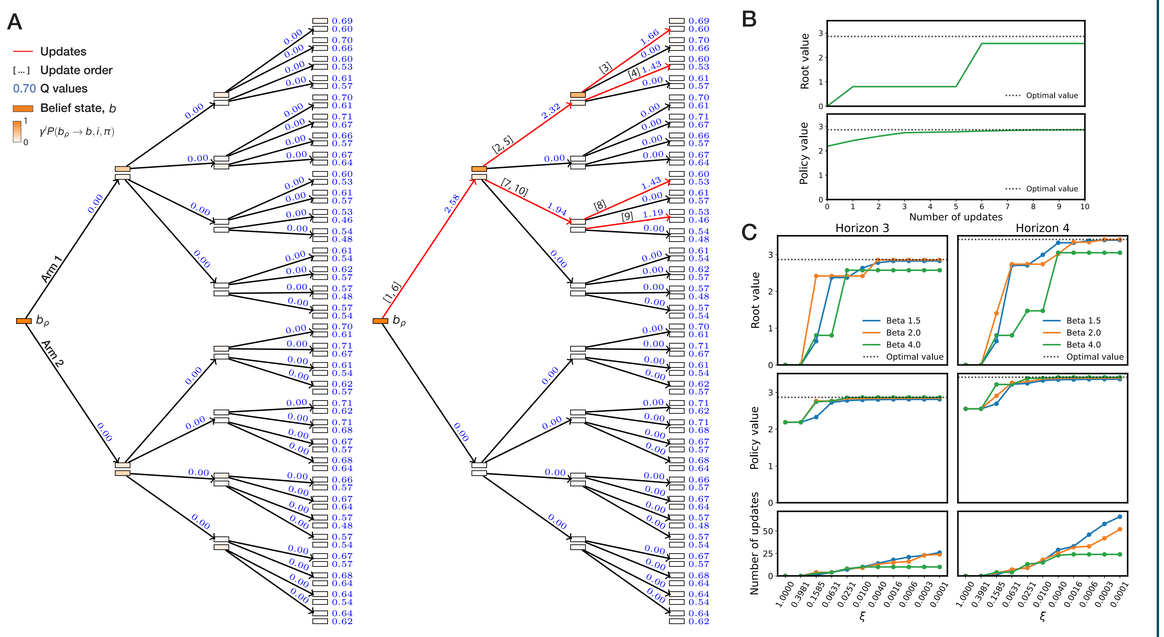
Although largely successful, the theory fails to account for sufficient exploration which is necessary for optimal behaviour

Bayes-adaptive policies optimally trade-off exploitation and exploration since they explicitly consider the agent's epistemic (model) uncertainty about the state of the world

In this work, we extend the theory of replay to partially observable domains to show that optimised replay does favour uncertain outcomes whenever there is a potential long-range benefit of exploration



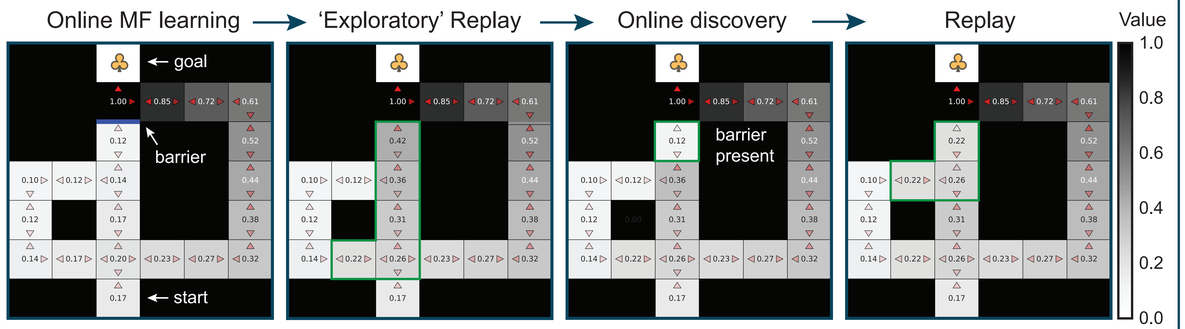
3 Prioritised sweeping in Bayesian bandits



First, we cosider a simple exploration/exploitation trade-off problem posed by finite-horizon planning in a 2-armed bandit

$$v_{\pi_{new}}(b) - v_{\pi_{old}}(b) = \gamma^i P(b \to b', i, \pi_{old}) \times \sum_a [\pi_{new}(a \mid b') - \pi_{old}(a \mid b')] q_{\pi_{new}}(b', a)$$

4 Offline replay drives directed exploration



We tested how well our agent performs exploration in Tolman's[4] detour maze

We treat the maze as a POMDP with unknown (latent) transition dynamics

The agent maintains a belief over the presense of one of the barriers from the original Tolman's experiment (blue line)

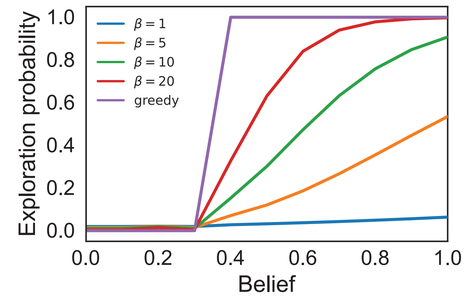
At the end of each trial, the belief is updated accroding to:

Forgetting

$$b_{t+1} = \begin{cases} \kappa\phi + (1-\kappa)b_t & \text{if the shortcut was not attempted} \\ 1 - \kappa(1-\phi) & \text{if the shortcut was successful} \\ \kappa\phi & \text{if the shortcut was unsuccessful} \end{cases}$$

Inference

and is set to what was actually experienced during the trial



For computing the optimal update order, we

$$v_{\pi_{new}}(z) - v_{\pi_{old}}(z) = \sum_{z' \in \mathcal{Z}} \sum_{i=0}^{\infty} \gamma^i P(z \to z', i, \pi_{old}) \times \sum_a [\pi_{new}(a \mid z') - \pi_{old}(a \mid z')] q_{\pi_{new}}(z', a)$$

5 Summary

6 References

1. Mattar MG, Daw ND. Prioritized memory access explains planning and hippocampal replay. Nat Neurosci 21, 1609–1617 (2018). https://doi.org/10.1038/s41593-018-0232-z
2. Sutton RS. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin 2.4, 160-163 (1991).
3. Moore AW, Atkeson CG. Prioritized sweeping: Reinforcement learning with less data and less time. Machine learning 13.1, 103-130 (1993)
4. Tolman add reference !!!!

GA and PD are funded by the Max Planck society. PD is also funded by the Humboldt Foundation