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 each potential individual replay update:

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

π_{old} Current policy [before replay update]
π_{new} New policy [implied by a replay update]

2 Problem

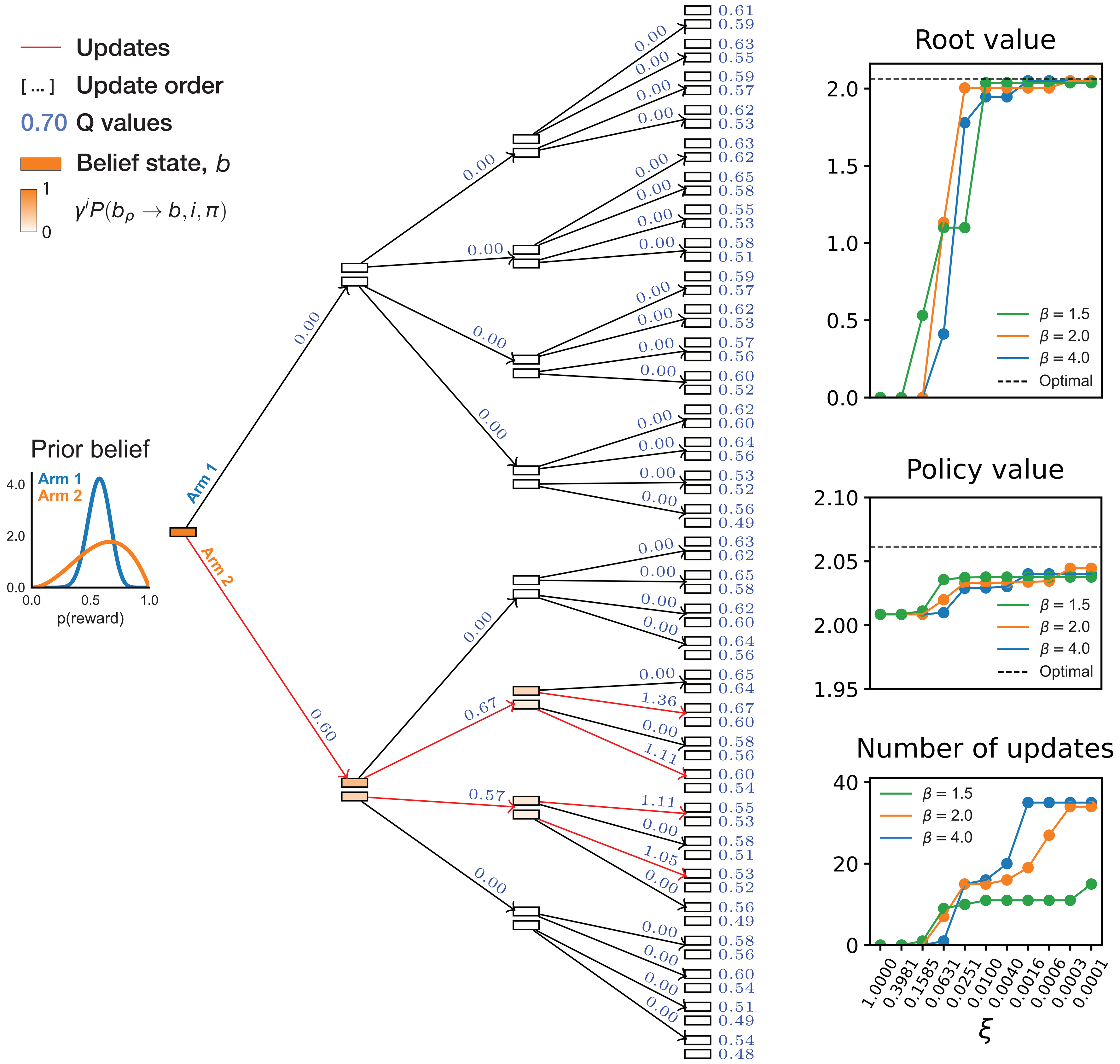
Although very successful, the theory does not extend to replay thatencourages optimistic exploration in the face of ignorance

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

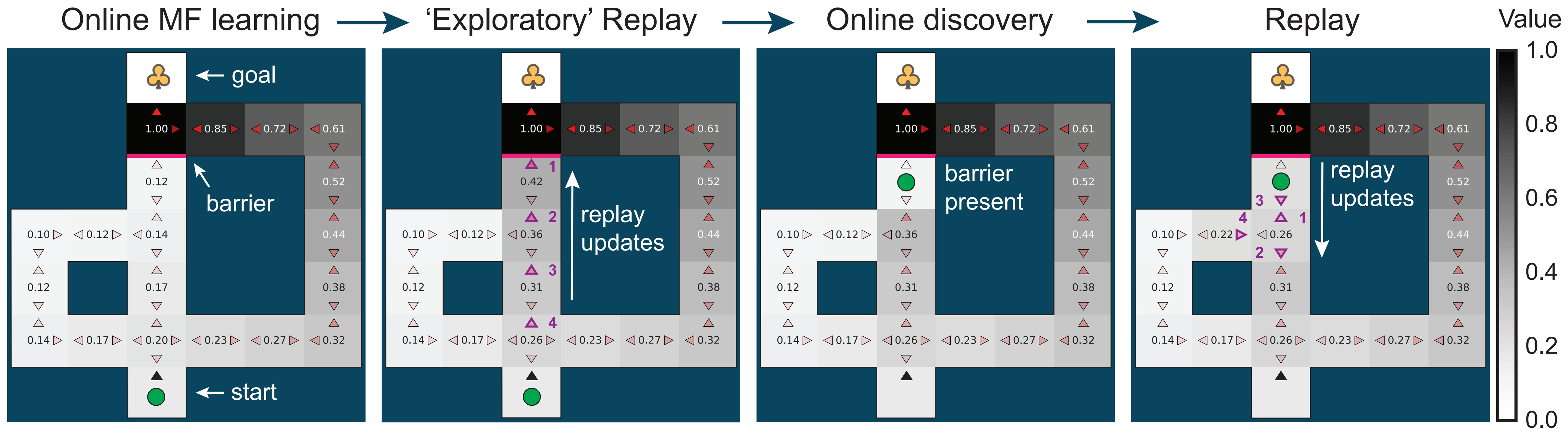
We consider optimised replay in belief MDPs where the following decomposition applies:

$$v_{\pi_{new}}(z) - v_{\pi_{old}}(z) = \sum_{z' \in \mathcal{Z}} \sum_{i=0}^{\infty} \gamma^i P(z \to z', i, \pi_{old}) \times \sum_a [\pi_{new}(a \mid z') - \pi_{old}(a \mid z')]\, q_{\pi_{new}}(z', a) \qquad \text{where } z = \langle b \in \mathcal{B}, s \in \mathcal{S} \rangle$$

3 Prioritised sweeping in Bayesian bandits



4 Offline replay drives directed exploration

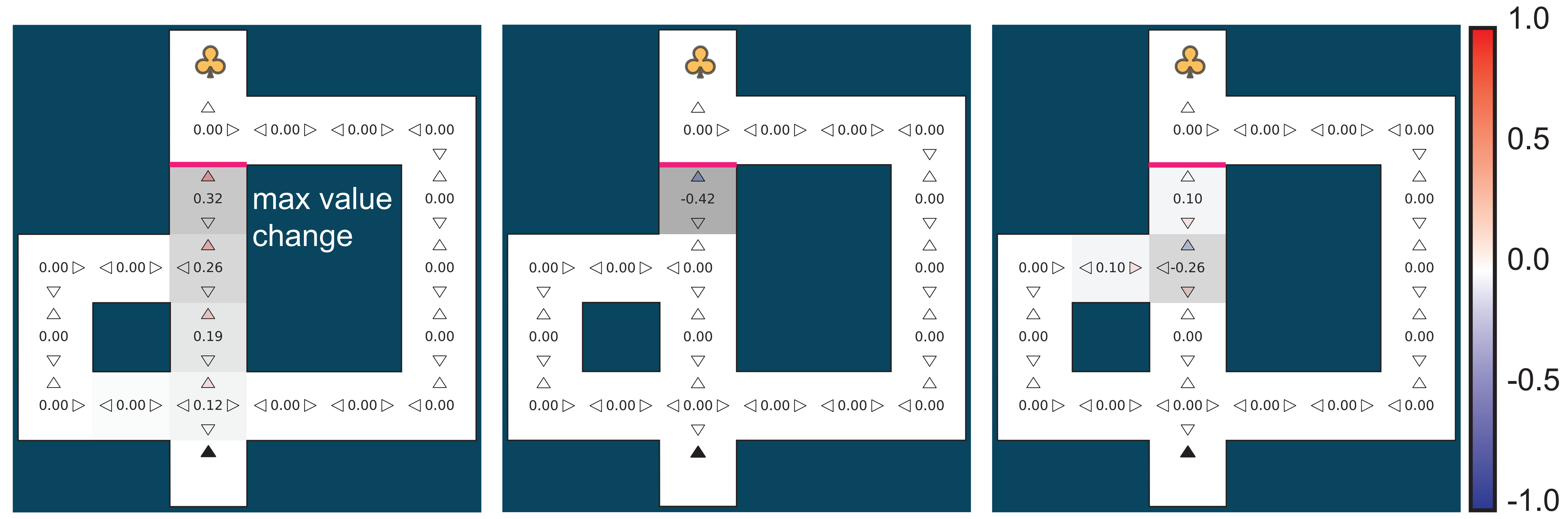


Value function changes occasioned by the above offline computations / behaviour

Tolman's[4] detour maze treated as a POMDP

The agent is uncertain about the barrier configuration

Offline replay in belief space can give rise to directed exploration



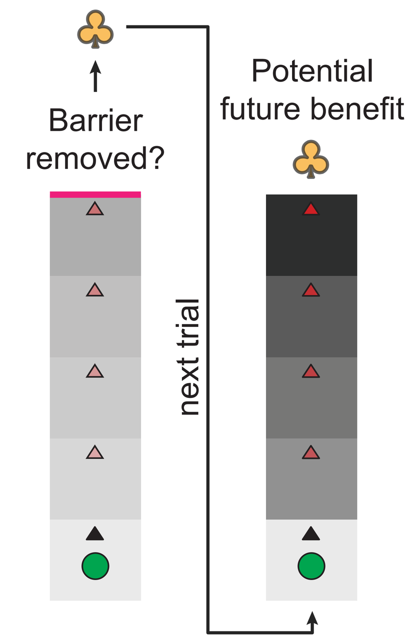
The agent maintains a belief over the presence of one of the barriers from the original Tolman's experiment

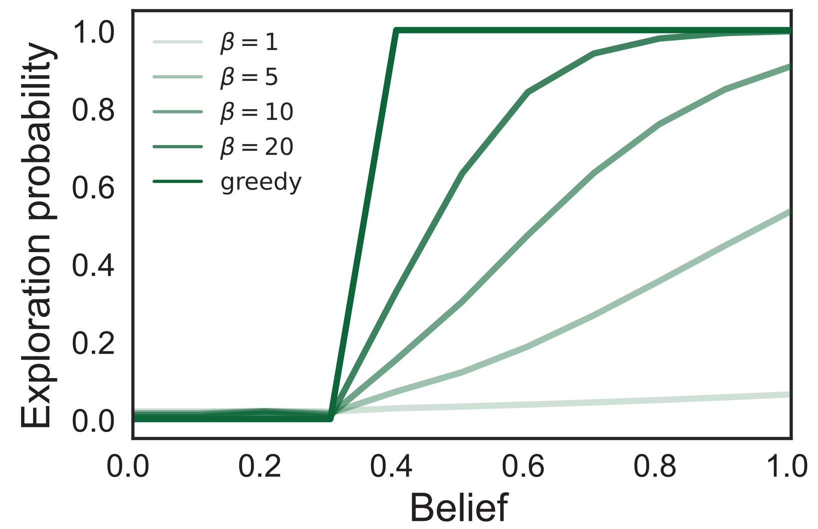
At the end of each trial, the belief is updated according to:

Inference

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

and is set to what was actually experienced during the trial





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