Sweeping Improvements to Exploration

Georgy Antonov[*]
Department of Computational Neuroscience
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany

Graduate Training Centre of Neuroscience
International Max Planck Research School
University of Tübingen
72076 Tübingen, Germany
georgy.antonov@tuebingen.mpg.de

Peter Dayan
Department of Computational Neuroscience
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany

University of Tübingen
72074 Tübingen, Germany
dayan@tue.mpg.de

Abstract

A modern synthesis of many studies examining hippocampal replay in decision-making tasks suggests that such patterns of behaviourally-relevant neural activity may support the sort of offline generative planning mechanisms such as DYNA that have been postulated in reinforcement learning (RL). A key observation in favour of this suggestion is the apparently close association between specific choices of replay experiences and both reward and the animal's policy – i.e., the decisions it subsequently makes; however, the rules that govern the selection of these experiences still remain poorly understood. A recent theory which is based on key optimising ideas from RL provides an astute normative account of this prioritisation, suggesting that replay experiences should be ordered according to their expected immediate impact on the value accrued by applying the newly changed policy. This theory closely matches experimental data from both rodent and human experiments; however, it focuses on exploitation to the exclusion of exploration, which limits its applicability. Here, we consider how offline, replay-like, planning mechanisms can contribute to information-seeking behaviour in the form of directed exploration by extending the theory to partially observable domains. We analyse the resulting exploratory replay choices in two cases: a stateless bandit with uncertainty about arm outcomes and a dynamic maze with a removable barrier which we model as a continual learning problem.

Keywords: Reinforcement learning, DYNA, planning, exploration, replay

Acknowledgements

GA and PD are funded by the Max Planck Society. PD is also funded by the Alexander von Humboldt Foundation. The authors thank Philipp Schwartenbeck and Christopher Gagne for their contributions to the early stages of this work.

[*]Corresponding author

1 Introduction

Mattar & Daw [1] insightfully combined two important methods from computational reinforcement learning [2], namely prioritized sweeping [3] and DYNA [4], with a central finding in the hippocampus of rodents [5] (since replicated in humans [6]) of the replay of patterns of activity associated with active behaviour during quiet wakefulness and sleep. They suggested that replay was a computationally efficient scheme (the prioritized sweeping) by which information from a generative model of the world could train a model-free (MF) policy (e.g., a Q-learning agent [7]) during offline periods (DYNA) such that the latter could be both reactively fast (by being MF) but also environmentally flexible (through model-based, or MB, instruction).

However, the original intention of DYNA was not to endow MF policies with MB properties, but rather to enable continual exploration. It did this by awarding recently unvisited state-action pairs with bonuses, and allowing offline MB training to propagate this information into a suitably exploratory policy. By contrast, Mattar & Daw [1] choose replays that are biased towards states that the agent's current policy already expects to visit, which corresponds to exploitation.

The optimal trade-off between exploration and exploitation given incomplete information comes from Bayes-adaptive policies that correctly account for the known uncertainty about the state of the world. In this work, we extend the ideas of optimal replay to partially observable domains, and study how it can thereby account for uncertainty-driven, reward-directed, information seeking.

2 Background

Gain and Need: By assessing the expected change in value (EVB) of a privileged start state s that would be caused by an update of an old policy, π_{old}, to a new one, π_{new}, Mattar & Daw [1] showed that the priority of an individual replay experience a_k at a potential state s_k should be jointly determined by two factors – Gain and Need:

$$v_{\pi_{\text{new}}}(s) - v_{\pi_{\text{old}}}(s) = \underbrace{\sum_{i=0}^{\infty} \gamma^i P(s \to s_k, i, \pi_{\text{old}})}_{\text{Need}} \times \underbrace{\sum_a \left[\pi_{\text{new}}(s_k, a) - \pi_{\text{old}}(s_k, a)\right] q_{\pi_{\text{new}}}(s_k, a)}_{\text{Gain}} =: \text{EVB}(s_k, a_k) \tag{1}$$

where $q_{\pi_{\text{new}}}(s_k, a)$ is the *true* value of performing action a at state s_k under the new (updated) policy, and $P(s \to s_k, i, \pi_{\text{old}})$ is the probability of making a transition from s to s_k in i steps, following policy π_{old}.

Gain associated with a particular replay update quantifies the expected *local* improvement in the agent's policy that would be engendered by that replay. This local measure of policy improvement is in fact very similar to the prioritized sweeping scheme proposed by Moore & Atkeson [3].

The crucial difference in Mattar & Daw, however, is the Need term, which is a more global measure of how often the agent expects to benefit from that local policy improvement in the future. Need is closely related to the successor representation [8] which quantifies the expected time-discounted future state occupancy of state s_k given the environmental transition dynamics and the agent's policy. Thus, under this scheme, an update at a state which the agent's current policy does not expect to visit often enough will not be considered suitably beneficial; this comes from the agent's inability to account for the changes to its policy and transition model as a result of the potential information gain.

BAMDPs: Bayes-Adaptive MDPs (BAMDPs) are a special case of partially observable MDPs (POMDPs) where the environmental transition dynamics are assumed to be unknown [9]. Importantly, in a BAMDP, agents transition through physical states, which are perfectly observable, as well as belief states, which represent the agent's accumulated knowledge about the environmental transition dynamics. Jointly, the agent's physical location and its belief state are referred to as information states, typically denoted as $z = \langle s \in \mathcal{S}, b \in \mathcal{B} \rangle$ – of which there is an infinitude.

For an agent that learns continually, each information state in a BAMDP can be visited at most once. This constitutes a major difference between planning in MDPs and BAMDPs – in the latter, it involves accounting for future learning opportunities which may improve one's decision quality. Moreover, revealing a piece of information in one information state typically affects knowledge in other information states – hence allowing broad generalisation. In spite of the incessant learning, each physical state should nonetheless have an accumulated Need in repetitive tasks such as mazes; and the individual belief states should also have different preferences according to the agent's policy – thus providing an opportunity for optimised and efficient planning. As a way of understanding exploratory replay in humans and other animals, we examine offline versions of planning in BAMDPs.

3 Results

Prioritised sweeping in Bayesian bandits

We start from the simple exploration/exploitation trade-off posed by finite horizon planning in a two-arm Beta-Bernoulli bandit. Such planning problems can be visualised as belief trees (Figure 1A) where each action (pulling an arm) can transition the agent to a number of different belief states (for Bernoulli bandits there are always two possibilities) according to their respective prior probabilities. Note that a full dynamic programming (DP) solution corresponds to backing up the values of terminal (leaf) nodes reached at the final horizon all the way to the root of the planning tree, weighted by their probabilities and assuming a greedy (optimal) policy at each decision step. Full DP solutions are computationally expensive, since the number of belief states in such planning trees explodes exponentially (the curse of expanding grid).

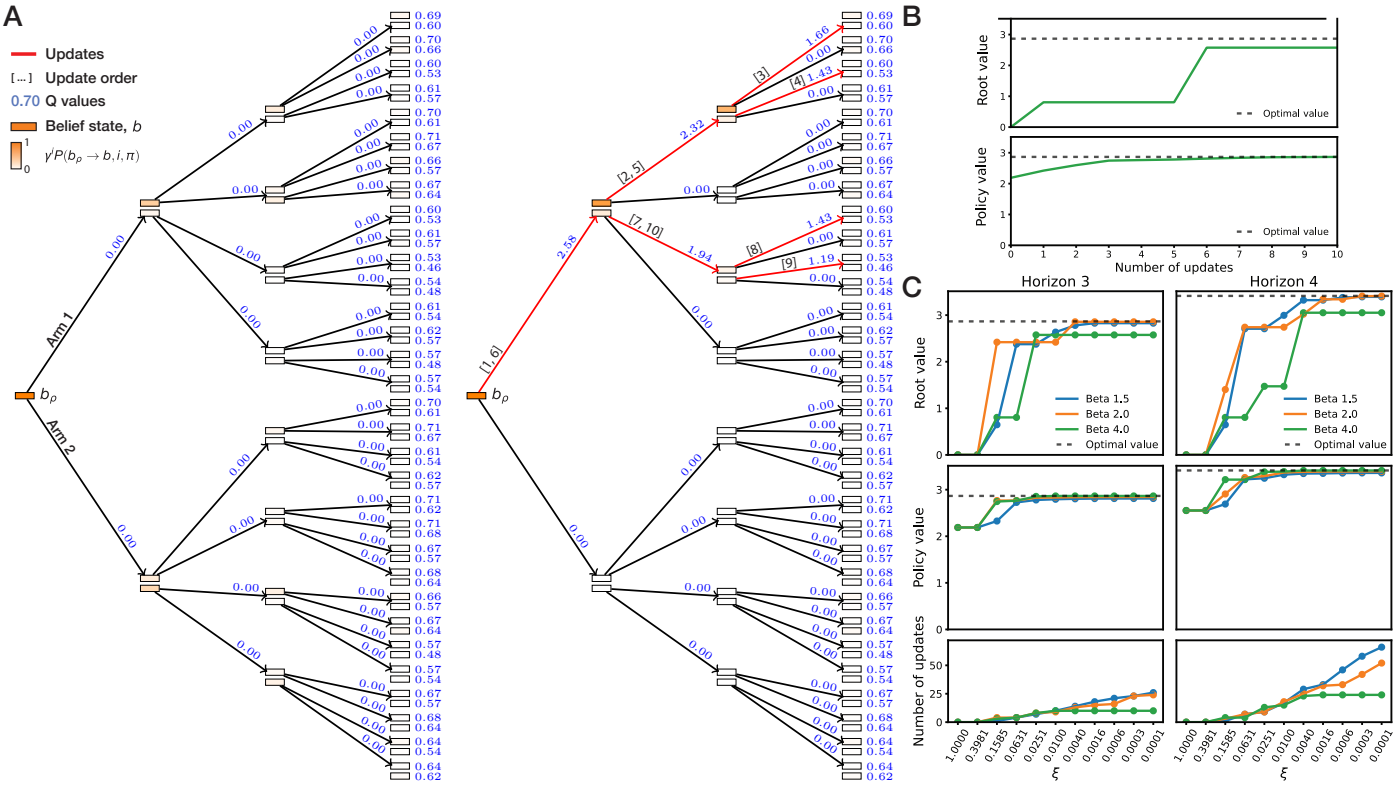

Figure 1: **Selective updates in belief trees.** (A) Left: belief tree built from the current belief state up to horizon 3. All belief states are depicted as orange rectangles with black outlines. The opacity of each belief state corresponds to the (discounted) probability with which the agent expects to reach that belief state. Black arrows depict the possible actions; top and bottom arrows from each belief state correspond to choosing arm 1 or 2 respectively. Similarly, among the paired belief states the ones on top are always a result of a successful outcome (received a reward) whereas the bottom ones are due to failed outcomes (received no reward). The root belief was initialised to Beta$(5, 1)$ for arm 1 and Beta$(2, 4)$ for arm 2. Blue numbers denote Q-values. Right: same tree as before but after 10 executed updates. The actions whose Q-values were updated are shown in red; the order of the executed updates is written in square brackets on top of the respective actions. (B) Upper: evolution of the root belief state value in the tree shown in (A) as a function of the 10 updates. Lower: same as above but for the true resulting policy value – that is, the value of the policy which was evaluated in the tree after 10 updates. (C) Top and middle: same as (B) but for varying horizons and inverse temperature parameters, as well as the EVB threshold, ξ. Bottom: number of executed updates as a function of the EVB threshold, ξ.

Behaving greedily during each successive value propagation means that at each decision point one of the two branches is always cut off, since the values of optimal actions are always propagated with probability 1. Thus, it is possible to optimise the selection of values for backward propagation. Inspired by the ideas of Moore & Atkeson [3] as well as Mattar & Daw [1], we consider the following prioritised sweeping rule to order the individual updates within the belief tree (the derivation is the same as in Mattar & Daw [1]):

$$\mathrm{EVB}(b, a_k) = \gamma^i P(b_\rho \to b, i, \pi_{\mathrm{old}}) \times \sum_a \left[\pi_{\mathrm{new}}(b, a) - \pi_{\mathrm{old}}(b, a)\right] q_{\pi_{\mathrm{new}}}(b, a) \tag{2}$$

The second term on the right-hand side is a belief state version of Mattar & Daw's Gain, and $P(b_\rho \rightarrow b, i, \pi_{\text{old}})$ is the probability that the agent reaches belief b from its current root belief b_ρ when following the current tree policy π_{old} in i steps (which is the horizon of that belief). We used $\gamma = 0.9$ in all subsequent simulations. As discussed above, the individual belief states are only encountered once – and hence this is a non-cumulative instance of Need.

The tree policy plays a critical role in determining the order of updates in our prioritisation scheme (just as it does in Monte-Carlo tree search [2]). Moreover, if one assumes a hybrid architecture (see below) where the root Q-values are initialised to the MF values, then it is important to include stochasticity to make sure that the optimal (according to the agent's belief) branches are assigned positive probabilities of enjoying replay. In all subsequent simulations we therefore use a softmax tree policy with an inverse temperature of 4.

The left tree in Figure 1A shows all the possible belief states the agent can reach up to horizon 3 from its current belief b_ρ at the root of the tree. The starting Q-values at each belief were assumed to be unknown, and hence initialised to 0; Q-values for the beliefs reached at the final horizon were initialised to the expected immediate reward according to those beliefs. The tree on the right of Figure 1A shows the update order specified by this algorithm, as well as the resulting Q-values at all the individual belief states. Figure 1B shows the evolution of the agent's estimate of the root value as a function of the number of updates it executed. Crucially, because distal beliefs can be updated, their contribution to the value at the root is not immediately obvious. In Figure 1B we therefore also show how the true value of the resulting policy evaluated in the tree changes as a function of each update. Note that the algorithm arrived at the near-optimal policy value within 10 updates, which is much fewer than a full DP procedure would have required.

An additional parameter in our algorithm, the EVB threshold ξ, controlled the minimal estimated benefit needed for each update to be executed. Figure 1C additionally shows how the root and policy values, as well as the number of updates the algorithm decides to execute, evolve as a function of EVB threshold.

Replay in the BAMDP

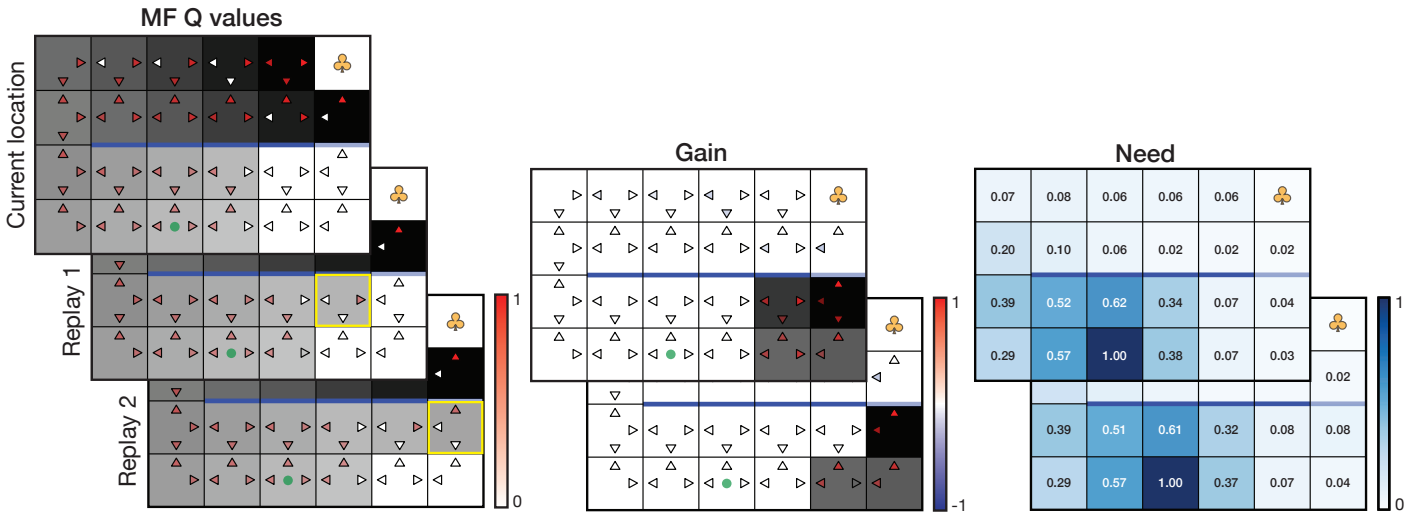

Figure 2: **Exploratory replay in a DYNA maze.** Top row: current state of knowledge of the agent. The arrows correspond to the actions available to the agent at each state; the colour intensity of the arrows corresponds to the respective model-free Q-values. The shading of each state additionally shows the *presumed* value of the best action available at each state. The green dot shows the current location of the agent in the maze, which also happened to be each trial's starting state. The yellow clover shows the goal state at which the agent receives a reward. The blue line corresponds to a barrier which the agent could not cross, except at the rightmost edge of the maze. The opacity of that particular barrier segment represents the agent's belief about whether it might be open or closed. Middle and bottom rows: two consecutive replays executed by the agent are highlighted by the yellow box in the left column. The middle column shows the Gain that the agent estimated for each individual replay update. Here the action colour intensities and state shadings show the magnitude of the (normalised) estimated Gain and the highest Gain available at each state respectively. Right: certainty-equivalent (normalised) Need computed by the agent according to its current belief about the transition model and the model-free policy.

Finally, we turn our attention to the maze case. We designed an environment reminiscent of the original DYNA maze shown in Figure 2. The agent was first allowed to learn the policy for going the long way towards the goal state; after it successfully has done so, the agent was informed that there is a 50% chance that the rightmost barrier might be open

– thus providing an opportunity for a potentially more direct route to the goal. This was operationalised by setting the agent's prior belief about the outcome of that transition to Beta$(1, 1)$.

To decide which replay updates to execute, the agent estimated the benefit of each potential update as for the root Q-values in the bandit belief tree example from Figure 1A. That is, each action's outcome at each state was planned up to a fixed horizon (in our simulations we used horizon 3), starting with the agent's current belief. However, because of the joint state-belief dynamics, the prioritisation within each state-action tree was determined by:

$$\mathrm{EVB}(z, a_k) = \gamma^i P(z_\rho \to z, i, \pi_{\mathrm{old}}) \times \sum_a \left[\pi_{\mathrm{new}}(z, a) - \pi_{\mathrm{old}}(z, a)\right] q_{\pi_{\mathrm{new}}}(z, a) \tag{3}$$

where $z_\rho = \langle s_\rho, b_\rho \rangle$ is the root state-action information state of each tree. The values of the different information states at each horizon were initialised to the agent's current MF Q-values; hence, this MB value estimation procedure can be thought of as a Bayesian prioritised sweep-until-habit process which optimally updates the values of information states.

For the updates at the root of each state-action tree, we used Mattar and Daw's evaluation from equation 1. This way, Need contained the agent's estimate of the expected future occupancy of each physical state conditioned on its current belief – which is a form of certainty-equivalence approximation. Gain, on the other hand, contained a fixed-horizon amount of potential learning (information gain) available at that state. The net effect of the fixed horizon in the maze is that even if, as in DYNA, there is continual change in the environment (e.g., from forgetting [10]), it suffices to consider this maze as a BAMDP rather than a full POMDP.

The top plot in Figure 2 shows the agent's current state of knowledge as well as its physical state (green dot) and belief state (transparency in blue of the rightmost barrier). In the middle row, we highlight the first replay that the agent decided to execute. Note that the state at which that replay update was chosen had very low Need (third column) because it deviated from the agent's current policy; however, the Gain (second column) that the agent estimated for that action was sufficiently high to result in an update (because the possibility of crossing the barrier was within the horizon reach). This replay, in turn, changed the agent's policy at that state, which resulted in an increased Need for the state immediately adjacent to the potential shortcut – and hence that action was updated next.

As a final remark, we would like to emphasise that that EVB quantity that we defined here for information states is still an approximation to an optimal exploration because of the use of MF heuristics. Moreover, it underestimates the benefit of generalisation of knowledge across information states because the expected value of updates at distal beliefs is computed with respect to the root of that tree – although many more beliefs can receive a potential benefit from those updates. The certainty-equivalent form of Need for the root updates is also a limitation, since it does not account for the learning that may happen on the way to the state at which a potential replay is considered. These assumptions and limitations are the targets for our future computational work. From an empirical viewpoint, it will be interesting to model tasks providing suggestive evidence for exploratory replay [11].

References

1. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience* **21,** 1609–1617 (2018).
2. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
3. Moore, A. W. & Atkeson, C. G. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning* **13,** 103–130 (1993).
4. Sutton, R. S. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin* **2,** 160–163 (1991).
5. Diba, K. & Buzsáki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience* **10,** 1241–1242 (2007).
6. Kurth-Nelson, Z., Economides, M., Dolan, R. J. & Dayan, P. Fast sequences of non-spatial state representations in humans. *Neuron* **91,** 194–204 (2016).
7. Watkins, C. J. C. H. *Learning from delayed rewards* PhD thesis (King's College, Cambridge, 1989).
8. Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Computation* **5,** 613–624 (1993).
9. Duff, M. O. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes* (University of Massachusetts Amherst, 2002).
10. Antonov, G., Gagne, C., Eldar, E. & Dayan, P. Optimism and pessimism in optimised replay. *PLOS Computational Biology* **18,** e1009634 (2022).
11. Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. Hippocampal place cells construct reward related sequences through unexplored space. *Elife* **4,** e06063 (2015).