Rational Arbitration of Hippocampal Replay

Mayank Agrawal (mayank.agrawal@princeton.edu), Marcelo G. Mattar (mmattar@princeton.edu), Nathaniel D. Daw (ndaw@princeton.edu), Jonathan D. Cohen (jdc@princeton.edu)
Princeton Neuroscience Institute and Department of Psychology
Princeton, NJ 08544

Abstract

It has recently been proposed that hippocampal replay can be explained in a reinforcement learning setting as the instantiation of the Dyna framework. This formulation lends itself to a dual-process model: an agent can choose to either act or replay at every time step. Here, we extend the proposed model by adding a controller that arbitrates between replaying and acting in order to maximize reward rate. That is, rather than give a fixed budget of replays to perform in both the starting and final state, we allow the agent to dynamically decide how much to replay in all states. The first result is that, in a Gridworld task, this algorithm is able to converge to the optimal policy faster. Second, by tracking the amount of replay selected per trial, we observe that there is only a narrow range of trials in which replay is beneficial. We propose this model as both a more efficient use of the Dyna framework as well as a normative model of how rational animal and human agents should replay.

Keywords: hippocampal replay; dual-process; control; reinforcement learning; rational analysis

Introduction

Dual-process models are ubiquitous in cognitive psychology and neuroscience. Often, they pit a fast, reflexive system versus a slow, deliberative system. This creates a higher-level form of a speed-accuracy tradeoff (in the choice between processes), and normative models of control aim to optimally decide the conditions under which each system should be used.

Recently, Mattar and Daw (2018) proposed a dual-process model incorporating hippocampal replay, the re-activation of place cells in rodents at rest. Using a reinforcement learning framework, they suggested that agents use replay in order to propagate reward information and thus receive the reward quicker. Their algorithm builds on the Dyna framework (Sutton, 1990), in which agents act with a model-free learner but are allowed to simulate experience using their knowledge of the world.

When Sutton first proposed the Dyna framework, he noted that optimal search control must address two questions, concerning *what* to simulate and *when* to simulate. The Mattar & Daw model offered a solution of *what* to simulate. Here, we tackle the second question: *when* (and, correspondingly, *how much*) to simulate. We show that a control mechanism able to adaptively determine when and how much to simulate outperforms the previous Mattar & Daw model. Furthermore, it allows us to model how the arbitration between acting and replaying changes over time, demonstrating that an agent does not need to replay after a certain amount of trials. Lastly, our results point to a potential relationship with paradigms in which agents must optimally trade off exploration vs. exploitation in order to maximize reward rate.

Background

Markov Decision Processes

Sequential decision problems are often modeled as a Markov decision process (MDP). An MDP $\mathcal{M}$ is a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{R}(s)$ is the reward received in state s, $\mathcal{P}(s,a,s')$ is the probability of transitioning to from state s to state s' using action a, and γ is the discount factor. A policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ is a function that determines with what probability an agent should perform action a when in state s, and the goal of the reinforcement learning agent is to identify the policy that maximizes its reward.

Model-Free Learning

There are two canonical types of reinforcement learning: model-free learning and model-based learning. Model-free algorithms do not require an agent to have any knowledge about the environment's transition structure. Rather, it keeps a table that maps state-action pairs to values and chooses how to act based on a simple lookup operation.

One of the most commonly used model-free algorithms is Q-learning (C. J. C. H. Watkins, 1989). In Q-learning, an agent keeps a table of all $Q(s,a)$ values and, after every state-action-state transition (s,a,s'), updates according to the rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(\mathcal{R}(s') + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

where α refers to the learning rate of the agent.

C. J. Watkins and Dayan (1992) demonstrated that Q-learning converges to the optimal policy π^* given sufficient experience. However, *sufficient* is the problem. If the goal is to maximize *reward rate*, then the question arises: how can Q-learning be sped up?

Dyna

Dyna is a framework proposed by Sutton (1990) in order to speed up model-free learning. In addition to acting with a model-free system, the agent can have a model of the environment that it uses to generate simulated experience and train the model-free system "offline:" the agent replays a $(s,a,s',\mathcal{R}(s'))$ tuple and then uses the standard Q-learning equation to update its Q-values.

Sutton (1990) demonstrated that these simulated experiences speed up the process of learning. However, in his initial

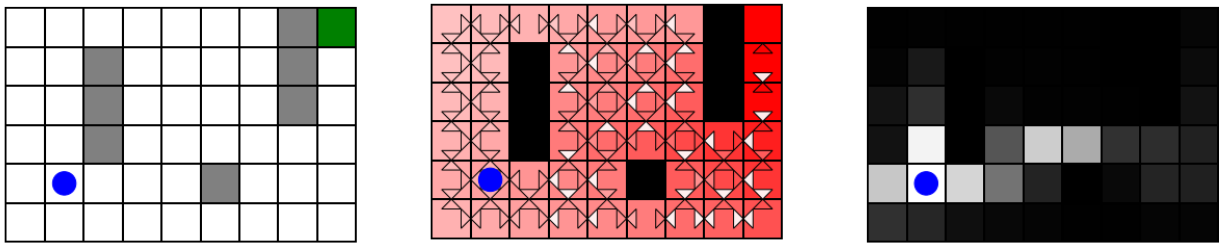

Figure 1: Example run in the Gridworld environment used to do model comparison. The left image shows the maze environment, the agent's location, and the reward location. The middle image shows the value of the states, where darker red refers to a higher value. The arrows colors reflect their Q-value. The right image shows to the estimated successor representation, where lighter shades designate states that are expected to be visited more in the future.

formulation of Dyna, experiences were simulated randomly. He noted that an improvement would be to add "search control;" that is, optimization of *what* and *when* to replay. Along these lines, research in reinforcement learning (Moore & Atkeson, 1993; Peng & Williams, 1993; Schaul, Quan, Antonoglou, & Silver, 2015) has generated heuristics for deciding what to replay. Recently, Mattar and Daw (2018) offered an algorithm to calculate the optimal experiences to replay. For every potential experience $(s_k, a_k, s'_k, \mathcal{R}(s'_k))$ to simulate, they computed the Expected Value of Backup (EVB):

$$\begin{aligned} EVB(s_k,a_k) &= \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t=s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t=s\right] \\ &= Gain(s_k,a_k)\times Need(s_k) \end{aligned}$$

where

$$Gain(s_k,a_k) = \sum_{a\in A} Q_{\pi_{\text{new}}}(s_k,a)\big(\pi_{\text{new}}(a|s_k) - \pi_{\text{old}}(a|s_k)\big)$$

and

$$Need(s_k) = \sum_{i=0}^{\infty}\gamma^i \delta_{s_{t+i},s_k}$$

The gain term is computed online while the need term is taken from the estimated successor representation (Dayan, 1993), $\widehat{M}$. Applying this algorithm to a Gridworld scenario, Mattar and Daw (2018) demonstrated that a reinforcement learning agent replaying the memories with the highest EVB exhibits similar qualitative patterns to place cell activations in rodents.

Rational Arbitration of Hippocampal Replay

Dual-process models also raise the question of *when* each system should be used (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011). Keramati et al. (2011) proposed that arbitration between a model-free and a model-based system should be done with respect to reward rate maximization. That is, given that the model-based system is more accurate but comes at the cost of time, it should only be used if the gains in accuracy outweigh the potential loss of reward in that amount of time.

Here, we apply a similar methodology for considering the value of replay in the model proposed by Mattar & Daw, to determine when it is worth engaging in replay. At every time step, the agent has the option to act or to replay, and we consider the cost of replaying instead of acting to be the opportunity cost of time (Kurzban, Duckworth, Kable, & Myers, 2013; Shenhav, Botvinick, & Cohen, 2013).

Let τ refer to the ratio between the time it takes to replay and the time it takes to act. One estimate of τ is 0.04, that comes from the speed of sharp wave ripples in the hippocampus; these occur at approximately 1,000 cm/s (Pfeiffer & Foster, 2013) as compared to the speed of running on a track, which is approximately 40 cm/s (Wikenheiser & Redish, 2015).

We extend the previous EVB and introduce EVB_C, that calculates the expected value of replay with an opportunity cost, by positing that the new expected future reward is additionally discounted by γ^τ:

$$\begin{aligned} EVB_C(s_k,a_k) &= \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=\tau}^{\infty}\gamma^i R_{t+i}|S_t=s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t=s\right] \\ &= \gamma^\tau \mathop{\mathbb{E}}_{\pi_{\text{new}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t=s\right] - \mathop{\mathbb{E}}_{\pi_{\text{old}}}\left[\sum_{i=0}^{\infty}\gamma^i R_{t+i}|S_t=s\right] \\ &= \widehat{M}(s,s_k)\sum_{a\in A} Q_{\pi_{\text{new}}}(s_k,a)\big(\gamma^\tau\pi_{\text{new}}(a|s_k) - \pi_{\text{old}}(a|s_k)\big) \end{aligned}$$

Now, let $EVB_C^* = \max_{(s_k,a_k)} EVB_C(s_k,a_k)$. If $EVB_C^* > 0$, the agent should replay the corresponding state-action-state experience. However, once the value of replaying falls below zero, it should act in accordance with its policy because replaying is now not worth the opportunity cost.

Methods

Like Mattar and Daw (2018), we use a 9×6 Gridworld environment with three sets of walls. The agent starts at $(1,3)$ and must gain the reward at $(9,6)$. A sample run of the environment is displayed in Figure 1.

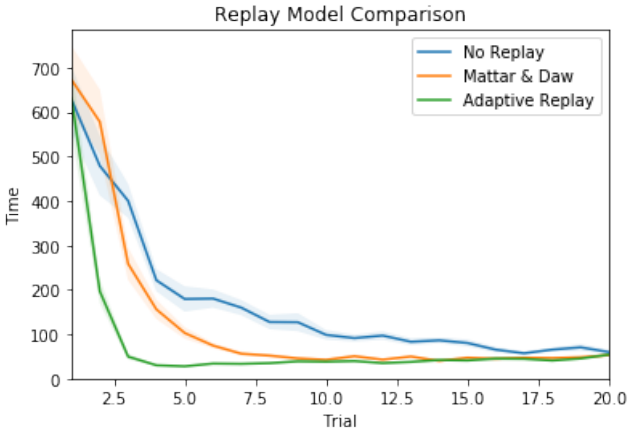

Figure 2: Time it takes for each algorithm to complete a given trial, which is calculated as the number of physical actions plus τ times the number of experiences replayed. Error shading indicates ± 1 SEM.

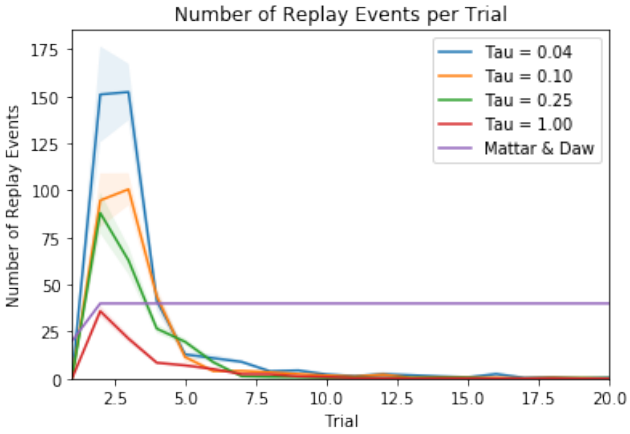

Figure 3: Number of experiences replayed every trial for the adaptive replay agent over a range of potential values of τ, compared with that of Mattar & Daw. Error shading indicates ± 1 SEM.

We set α and γ to values of 0.9, and the agent learned the successor representation using the standard temporal difference learning algorithm (with α and γ values again set to 0.9). The agent acted with a softmax policy, i.e.

$$\pi(a_t|s_t) = \frac{e^{\beta Q(s_t,a_t)}}{\sum_a e^{\beta Q(s_t,a)}}$$

in which β is the temperate parameter. For our simulations, we let $\beta = 5$.

Lastly, our implementation differs from the Mattar & Daw model in two ways. First, we restrict the agent to be able to simulate only previous experiences rather than all possible experiences. Second, the agent automatically restarts the maze after receiving the reward, and thus does not have the opportunity to replay post-reward in the same trial.

We compare this 'adaptive replay' model to (1) the Mattar & Daw model that replays twenty times at both the beginning and end of each trial, as well as (2) a model without any replay. In this simulation, we use $\tau = 0.04$.

Results

In Figure 2, we plot the amount of time it takes for each agent to find the reward during every trial. We conducted fifty simulations of each agent partaking in twenty trials apiece. The amount of time is calculated as the number of actions (i.e. moving up, down, left, right) in the trial plus τ times the number of replay events in the trial. We see that the Mattar & Daw model, which has forty replay steps per trial (except for the first trial, in which it has twenty), converges significantly faster than a model without replay. The adaptive replay model is even faster, converging to a stable policy between trials three and four. However, it should be noted that this faster convergence is a product of more replays, and if the fixed budget of the Mattar & Daw agent was set to the maximum replay amount by the 'adaptive replay' agent, that model may perform better than it currently is during the earlier trials.

To gain insight into the difference between these models, and the contribution made by the current one, we plot the amount of replay per trial for the adaptive replay agent in Figure 3. We also show the results of simulations using multiple values of τ, to determine its influence on replay frequency. Figure 3 demonstrates that, regardless of the value, there is consistently little replay at the beginning, followed by a rapid increase until it reaches a peak, and then a decrease as model-free learning converges to the optimal policy.

This behavior resembles an explore-exploit tradeoff (Wilson, Geana, White, Ludvig, & Cohen, 2014), as well as a switch to habitual (i.e. model-free) decision-making (Dolan & Dayan, 2013). The initial lack of replay presumably reflects the value of exploring the environment and acquiring experience that can be used later for replay. That is, when the agent has relatively uninformative Q-values and little experience, replay is uninformative. However, once the agent has gained sufficient experience to have informative Q-values, it prioritizes replay over acting as way of propagating this information quickly. This corresponds to the trials in which the replay is near its peak. As replay uses this information to inform the agent's Q-values, the need for replay diminishes. Eventually, the agent develops a sufficiently good model-free policy that diminishes the opportunity and thus favors action over replay. In sum, by maximizing reward rate, the adaptive replay model captures both the change from exploration to exploitation, as well as the transfer from primarily replaying to primarily acting in the exploitation stage.

Discussion

The strength of the Mattar & Daw model was not just faster reinforcement learning, but also the close match of its behav-

ior to that observed in rodents. Similarly, the extension we propose here may help explain additional empirical findings. Foster (2017) hypothesized that overtraining was the reason scientists took much longer to discover awake replay after the discovery of sleep replay, and that hypothesis has been supported by experiments demonstrating that replay occurs more in novel environments than familiar environments (Foster & Wilson, 2006; Diba & Buzsáki, 2007). Furthermore, it has also been suggested that replay increases with initial exposure (Buhry, Azizi, & Cheng, 2011). The results shown in Figure 3 support both of these claims.

Our formulation also leads to an interesting observation: in a given trial, replay has a lower opportunity cost at distances far from the reward, because the expected future discounted reward is lower. This can be tested in future empirical work, that tests: (1) the effect of a state's expected reward on the amount of replay, and, correspondingly, (2) the change in total amount of replay from novelty to habituation. Both the Mattar & Daw model and the adaptive replay model proposed here make the unrealistic assumption that the value of replay can be computed precisely before actually replaying. Discrepancies between empirical results and these models may suggest heuristics agents are using in order to estimate these values.

Conclusion

A rational agent seeks to maximize reward rate, not just reward. In this paper, we added a rational arbiter to the Mattar & Daw model, that that sought to maximize reward rate by determining whether to act or replay. We found that such an agent exhibited three phases during learning of a new task: an exploration phase in which it acquires experiences and builds up informative Q-values; a deliberative exploitation phase in which it uses these Q-values to replay; and a habitual exploitation phase in which it is able to act quickly based on mature Q-values. The model makes novel predictions that could be tested in future empirical work.

The model also raises questions about related theoretical assumptions. For example, it updated its successor representation only when an actual action took place, but not when a replay step was simulated. If replay is equivalent to another action, should it also affect the successor representation? If so, this raises the question of how the successor representation can be scaled to domains in which actions can take different amounts of time. Furthermore, the current model used a pre-specified value for τ, the ratio between the time required for a single replay and the time require for an overt action. However, it seems reasonable to assume that τ may vary for different environments, different agents, or even the same agent in different states. Thus, an interesting direction for future pursuit is to consider how an agents might estimate τ, and adaptively adjust in different conditions.

References

Buhry, L., Azizi, A. H., & Cheng, S. (2011). Reactivation, replay, and preplay: how it might all fit together. *Neural plasticity*, *2011*.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature neuroscience*, *10*(10), 1241.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Foster, D. J. (2017). Replay comes of age. *Annual review of neuroscience*, *40*, 581–602.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5), e1002055.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine learning*, *13*(1), 103–130.

Peng, J., & Williams, R. J. (1993). Efficient learning and planning within the dyna framework. *Adaptive Behavior*, *1*(4), 437–454.

Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, *497*(7447), 74.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. *arXiv preprint arXiv:1511.05952*.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990* (pp. 216–224). Elsevier.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Wikenheiser, A. M., & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nature neuroscience*, *18*(2), 289.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.