Adaptive patch foraging in deep reinforcement learning agents

Nathan J. Wispinski[*]
University of Alberta
Edmonton, Alberta, Canada
nathan3@ualberta.ca

Andrew Butcher
DeepMind
Edmonton, Alberta, Canada

Craig S. Chapman
University of Alberta
Edmonton, Alberta, Canada

Matthew M. Botvinick
DeepMind
London, UK

Patrick M. Pilarski
DeepMind & University of Alberta
Edmonton, Alberta, Canada

Abstract

When to explore and when to exploit is a fundamental decision problem that all biological agents must face. One ecological explore-exploit problem, patch foraging, provides a touchstone for artificial intelligence where biological intelligence is successful, adaptive, and sometimes optimal. Here we show deep reinforcement learning agents that can successfully and adaptively forage in a patchy three-dimensional environment. Agents learn to tradeoff exploration and the exploitation of patches, and strike this balance differently in scarce and plentiful environments similar to biological foragers. However, these agents tend to overstay in patches relative to the optimal solution from the marginal value theorem in behavioural ecology, suggesting potential key differences in how artificial and biological agents make tradeoffs during foraging.

Keywords: Foraging
Animal behaviour
Neuroscience
Reinforcement learning
Deep learning

Acknowledgements

We are deeply indebted to our DeepMind colleagues Leslie Acker, Andrew Bolt, Michael Bowling, Dylan Brenneis, Adrian Collister, Elnaz Davoodi, Richard Everett, Arne Olav Hallingstad, Nik Hemmings, Edward Hughes, Michael Johanson, Marlos Machado, Kory Mathewson, Drew Purves, Kimberly Stachenfeld, Richard Sutton, Jane Wang, and Alexander Zacherl for their support, suggestions, and insight regarding this work.

[*]Corresponding author. This work was conducted at DeepMind, with collaboration from the University of Alberta.

1 Foraging in biological and artificial agents

When to explore and when to exploit is a fundamental decision problem that all biological agents must face. Foraging at its core is one such explore-exploit problem, where biological agents must tradeoff the exploitation of resources they currently have access to with the exploration for more resources. In patch foraging theory, spatial patches are frequently modeled as exponentially decaying in resources, with areas outside of patches as having no resources (Charnov, 1976). Agents are faced with a fundamental decision about when to cease foraging in a depleting patch in order to begin travelling some distance to a richer patch. Research has shown that many animals are adaptive patch foragers in this context, intelligently staying in patches for longer when the environment is resource scarce, and staying a shorter time in patches when the environment is resource rich (Cowie, 1977; Hayden, Pearson, & Platt, 2011; Krebs, Ryan, & Charnov, 1974).

Foraging is so important to the survival of biological agents that theorists argue that the foraging behaviour of animals should not only be adaptive, but should approach optimal in natural environments because of strong selective pressures (Stephens & Krebs, 2019). In patch foraging, the marginal value theorem (MVT) offers a solution to optimal patch foraging behaviour (Charnov, 1976). In short, the MVT states that the optimal solution is to cease foraging within a patch and search for a new patch when the reward rate of the current patch drops below the average reward rate of the environment. Many animals, including humans, have been shown to behave optimally in patch foraging tasks in the wild and in the laboratory. For example, human mushroom foragers (Pacheco-Cobos et al., 2019), non-human primates (Hayden et al., 2011), and birds, fish, and bees (Cowie, 1977; Krebs et al., 1974; Stephens & Krebs, 2019) have all shown to behave consistent with the MVT solution of optimal patch foraging.

Many computational models of patch foraging are agent-based models with fixed decision rules (Tang & Bennett, 2010), although recent work has involved the use of tabular reinforcement learning models (Constantino & Daw, 2015; Goldshtein et al., 2020; Miller, Ringelman, Eadie, & Schank, 2017; Morimoto, 2019). Neural networks have also displayed foraging behaviour in ecological tasks such as patch selection (Coleman, Brown, Levine, & Mellgren, 2005; Montague, Dayan, Person, & Sejnowski, 1995; Niv, Joel, Meilijson, & Ruppin, 2002). Foraging behaviour has also been shown in deep reinforcement learning agents searching environments for rewarding collectibles like apples while avoiding obstacles and/or enemies (Lin, 1991; Platanios, Saparov, & Mitchell, 2020). However, these deep reinforcement learning environments often significantly differ from those in theoretical and experimental ecological research.

Here, we first ask if deep reinforcement learning agents can learn to forage in a 3D patch foraging environment inspired by experiments from behavioural ecology. Next we ask whether these agents forage intelligently—adapting their behaviour to the environment in which they find themselves. Finally, we investigate if agent foraging behaviour in these environments approaches the gold standard—the optimal solution determined by the marginal value theorem (Charnov, 1976). This paper provides the first investigation of deep reinforcement learning agents in an ecological patch foraging task, and adds to a literature suggesting that current reinforcement learning methods may fall short of foraging solutions in biological agents (Constantino & Daw, 2015; Miller et al., 2017). These experiments are described not as a performance benchmark for artificial agents, but rather as an empirical investigation into the emergence of complex patch foraging behaviour, and the potential limits of discounted reinforcement learning algorithms in ecological environments.

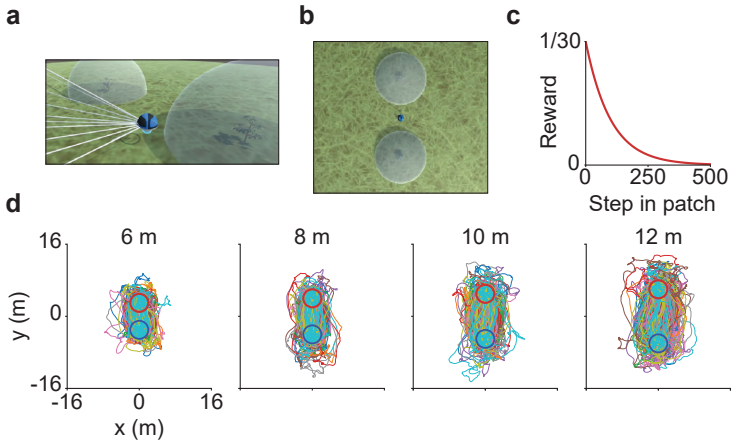

Figure 1: Task. **a)** Mock-up of the 3D foraging environment and agent with LIDAR rays. **b)** Overhead view. An agent starts each episode between two equidistant patches. **c)** The agent receives exponentially decreasing reward on every step it is within a patch. When the agent enters one patch, the opposite patch is refreshed to its starting reward state. **d)** Overhead spatial trajectories of a representative trained agent in each evaluation environment.

2 Experiments

Environment. A continuous 2D environment was selected to approximate the rich sensorimotor experience involved in ecological foraging experiments, as well as for future extensions into multi-patch foraging. The environment consisted of a 32 x 32 m flat world with two patches (i.e., half spheres) equidistant from the center of the world (Figure 1a; see Cultural General Intelligence Team et al., 2022). Patches always had a diameter of 4 m. Agents started each episode at the middle of the world, facing perpendicular to the direction of the patches. Each episode terminated after 3600 steps. Agents received a reward of zero on each step they were outside of both patches. When an agent was within a patch, it received reward according to the exponentially decaying function, $r(n) = N_0 e^{-\lambda n}$, where n is the number of non-consecutive steps the agent has been inside a patch without being inside the alternative patch. In this way, as soon as an agent entered a patch, the alternative patch was refreshed to its initial reward state (i.e., $n = 0$). As such, agents are faced with a decision about how long to deplete the current patch before traveling toward a newly-refreshed patch. For all experiments, the initial patch reward, N_0, was set to $1/30$, and the patch reward decay rate, λ, was set to 0.01 (Figure 1c). The surface color of each patch changed proportional to the reward state of the patch in RGB space. Patches changed color from white (i.e., [1, 1, 1]) to black (i.e., [0, 0, 0]) following the function, $r(n)/N_0$. In this way, agents had access to the instantaneous reward rate of the patch through patch color, rather than having to estimate patch reward rate by estimating the decay function and keeping track of steps spent within a patch.

Agents. Agents had a LIDAR-based observation space—a common sensory modality for physical robots (Malavazi, Guyonneau, Fasquel, Lagrange, & Mercier, 2018), and agents in other simulated environments (e.g., Baker et al., 2019; Cultural General Intelligence Team et al., 2022). The current agents had eight horizontal LIDAR rays evenly spaced throughout 90°, and 3 vertical rays evenly spaced throughout 60°, extending from the front center of the agent. Each LIDAR ray coded for distance (max distance 128 m; normalized), one-hot encoded object type, and RGB colour of the first object it intersected with. LIDAR inputs were convolved (24 output channels, 2x2 kernel shape), before they were concatenated with the reward and action taken on the previous step. These values were then passed through a MLP (3 layers of 128, 256, and 256 units) and a LSTM layer (256 units). Finally, LSTM outputs were passed to an actor and a critic network head. Agents were given a continuous action space that included strafing forward/backward and left/right, looking up/down and left/right, jump/crouch, and a grasp and use action that were not relevant for this environment. Actions were taken by sampling from Gaussians parameterized by the policy network head output for each action dimension. Agents were trained in a distributed manner, each interacting with 16 environments in parallel. Experience was saved in a buffer and agent parameter updates were accomplished via the MPO learning algorithm (Abdolmaleki et al., 2018). A neural network was chosen over tabular methods because of the complexity of the environment, and the potential for future comparisons of internal network dynamics to neural recordings in biological agents.

Three agents were trained in each of four discount rate treatments ($N = 12$), selected on the basis of MVT simulations (Figure 3d). Agents were each initialized with a different random seed, and trained for $12e^7$ steps using the Adam optimizer (Kingma & Ba, 2014) and a learning rate of $3e^{-4}$. On each training episode, patch distance was drawn from a random uniform distribution between 5 m and 12 m, and held constant for each episode. Trained agents were evaluated on 50 episodes of each evaluation patch distance (i.e., 6, 8, 10, and 12 m). If an agent was within a patch at the end of an evaluation episode, this final patch encounter was rejected from all analyses, as no distinct patch leave behaviour could be verified to determine the total steps in this patch.

3 Results

Trained agents displayed behaviour consistent with successful patch foraging—agents learned to leave patches before they were fully depleted of reward (mean leaving step = 121.7), and traveled for several steps without reward in order to reach a refreshed patch (mean travel steps between patches = 57.7). Agents also achieved a higher score on episodes where patches were closer together ($b = -5.82 \pm 0.21$, $p = 3.96x10^{-166}$).

In patch foraging, it is not only important to balance the exploitation of patches with exploration to find new patches, but also to intelligently adapt this balance when in more plentiful or more scarce environments. This adaptive behaviour is

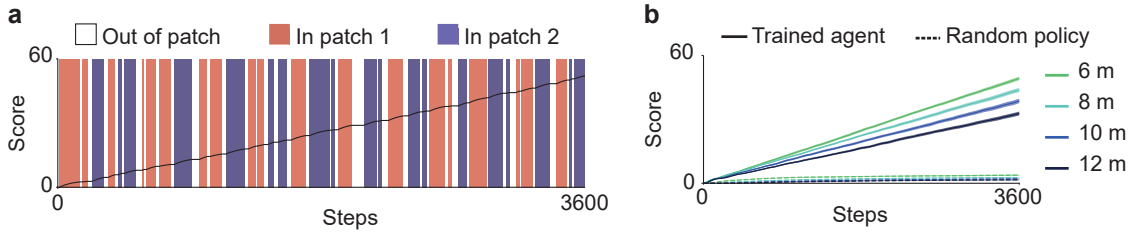

Figure 2: Performance. **a)** Agent behaviour from a representative evaluation episode. Shaded regions define when the agent is outside of any patch (white), inside patch 1 (red), or inside patch 2 (blue). **b)** Episode score for a trained agent (solid lines), and a random agent (dashed lines) in each evaluation environment. Shaded regions denote standard errors.

present in the current agents; trained agents adapted their patch leaving times to the environment, leaving patches later when travel distance is higher ($b = 9.60 \pm 0.87$, $p = 4.03x10^{-28}$; Figure 3a).

Above we show that trained agents are able to successfully forage, and intelligently adapt their foraging behaviour to the environment in accordance with patch foraging theory (Charnov, 1976; Stephens & Krebs, 2019), and animal behaviour (e.g., Cowie, 1977). However, do these agents adapt optimally according to the marginal value theorem (MVT), like many animals (Stephens & Krebs, 2019)? As stated above, the MVT provides a simple rule for when to leave patches optimally (Charnov, 1976). That is, an agent should leave a patch when the reward rate of the patch drops below the average reward rate of the environment. For each agent and evaluation environment (e.g., 6 m), we can estimate the average reward rate of the environment by calculating the average reward per step for each evaluation episode. Over all evaluation episodes, this provides an estimate of the optimal patch leaving step (Figure 3c). Comparing the difference between average observed and optimal patch residence times, agents tend to overstay in patches relative to the optimal solution, $t(11) = 5.60$, $p = 1.60x10^{-4}$ (Figure 3e). Bonferroni-corrected t-tests show that agents significantly overstayed relative to the MVT solution in all evaluation environments ($ps < 0.0015$), except for 12 m ($p = 0.047$).

The current agents however use temporal discounting methods, which exponentially diminish rewards in the future. Given that agents are effectively asked to compare the values between the current patch reward on the next step relative to a refreshed patch reward after several travel steps, temporal discounting encourages longer patch residence times. The difference between observed and optimal behaviour is modulated by the temporal discounting rate, where agents trained with higher temporal discounting rates tend to behave closer to optimal ($b = -2784.01 \pm 992.19$, $p = 0.019$; Figure 3f). Are agents then optimal after accounting for temporal discounting rates in the MVT solution? We accounted for the temporal discounting rate by simulating individual stay and leave decisions at many patch residence steps. Agents could either stay for an additional step of reward before leaving a patch, or immediately leave the patch, where the subsequent 5000 steps were simulated as alternating between a fixed number of steps in a patch and a fixed number of steps traveling between patches. Over a grid of fixed subsequent patch and travel steps, the difference in the discounted return (sum of discounted rewards) between each stay/leave decision provided an indifference curve, where the 5000-step discounted return was equal for staying relative to leaving. Where this stay/leave indifference step matched the fixed patch steps provided an approximation of an average patch time where the value of leaving is about to exceed the value of staying. After accounting for each agent's temporal discounting rate in the MVT (Figure 3d), agent patch residence times were closer to the optimal solution (Figure 3h). Bonferroni-corrected t-tests show that agents significantly overstayed in the 6 and 8 m evaluation environments ($ps < 0.0067$), understayed in the 12 m ($p = 0.0023$), and were not significantly different from optimal in the 10 m environment ($p = 0.30$).

4 Discussion

Here we tested deep reinforcement learning agents in a foundational decision problem facing biological agents—patch foraging. We find that these agents successfully learn to forage in a 3D patch foraging environment. Further, these agents intelligently adapt their foraging behaviour to the resource richness of the environment in a pattern similar to many biological agents (Stephens & Krebs, 2019).

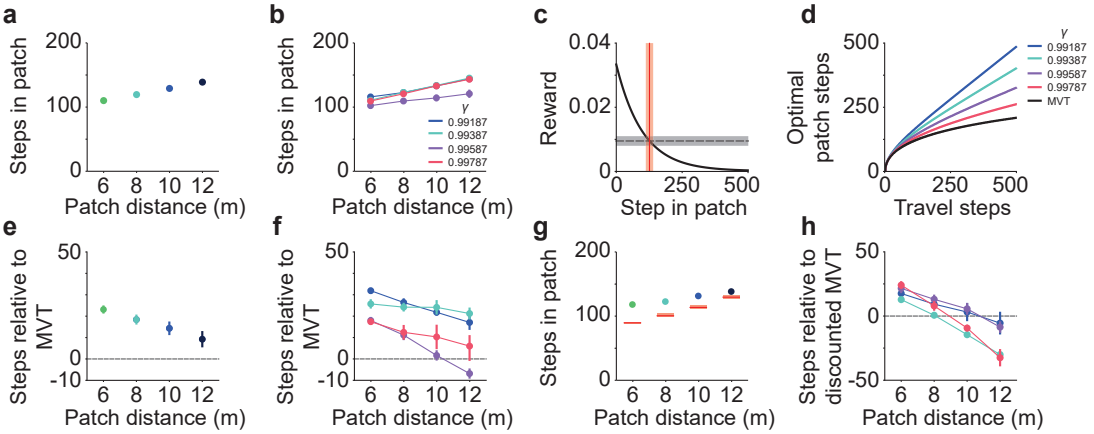

Figure 3: Patch residence times. **a)** Average of all agents. **b)** Average of agents grouped by discount rate. **c)** Graphical model of the MVT solution. Where the patch reward rate (solid black line) intersects the observed average reward rate of the environment as determined by the agent's behaviour (dashed horizontal black line), determines the MVT optimal average patch leaving time (solid vertical red line). **d)** Patch leaving times prescribed by the MVT (black), and simulation results for the MVT considering discount rates. **e)** Mean difference between the observed and optimal patch residence time for all agents, and **f)** agents grouped by discount rate. **g)** Representative single trained agent patch residence times (dots) against the MVT solution (red lines). **h)** Mean difference between the observed and discounted MVT patch residence time for agents grouped by discount rate. All vertical lines and/or shaded regions denote standard errors.

Many animals, including humans in the wild (Pacheco-Cobos et al., 2019), have been shown to be optimal patch foragers. The deep reinforcement learning agents investigated here are adaptive, yet sub-optimal foragers. These results are similar to those from humans in computerized patch foraging tasks, where they tend to overstay relative to the MVT solution (Constantino & Daw, 2015). We find that agents trained with higher temporal discounting rates tend to display patch leaving times closer to the MVT solution (Figure 3f). Further, in a human behavioural study, Constantino and Daw (2015) found that reinforcement learning methods that estimate cumulative long-term discounted rewards are a poor fit for human foraging behaviour in a lab setting, relative to methods which estimate the average reward per step. Overall, these findings suggest potential key differences in how many artificial and biological agents make tradeoffs during foraging, which may be reconciled with further work on average reward reinforcement learning (Sutton & Barto, 2018).

Biologists and behavioural ecologists may benefit from reinforcement learning approaches to agent-based modelling (Frankenhuis, Panchanathan, & Barto, 2019). Further, tasks from behavioural ecology, such as patch foraging, provide insights into fundamental decision problems facing intelligent biological agents. At its core, foraging is an explore-exploit tradeoff, and taking cues from how biological agents (often optimally) solve this dilemma may provide novel methods for artificial agents to do the same.

References

Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., & Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528*.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Coleman, S. L., Brown, V. R., Levine, D. S., & Mellgren, R. L. (2005). A neural network model of foraging decisions made under predation risk. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(4), 434–451.

Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(4), 837–853.

Cowie, R. J. (1977). Optimal foraging in great tits (Parus major). *Nature*, *268*(5616), 137–139.

Cultural General Intelligence Team, Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., … Zhang, L. M. (2022). Learning robust real-time cultural transmission without human data. *arXiv preprint arxiv.2203.00715*.

Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2019). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes*, *161*, 94–100.

Goldshtein, A., Handel, M., Eitan, O., Bonstein, A., Shaler, T., Collet, S., … others (2020). Reinforcement learning enables resource partitioning in foraging bats. *Current Biology*, *30*(20), 4096–4102.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*(7), 933–939.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behaviour*, *22*, 953–IN3.

Lin, L. J. (1991). Self-improvement based on reinforcement learning, planning and teaching. In *Proceedings of the Eighth International Conference on Machine Learning* (p. 323–327).

Malavazi, F. B., Guyonneau, R., Fasquel, J.-B., Lagrange, S., & Mercier, F. (2018). Lidar-only based navigation algorithm for an autonomous agricultural robot. *Computers and Electronics in Agriculture*, *154*, 71–79.

Miller, M. L., Ringelman, K. M., Eadie, J. M., & Schank, J. C. (2017). Time to fly: A comparison of marginal value theorem approximations in an agent-based model of foraging waterfowl. *Ecological Modelling*, *351*, 77–86.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, *377*(6551), 725–728.

Morimoto, J. (2019). Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data. *Journal of Theoretical Biology*, *467*, 48–56.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior*.

Pacheco-Cobos, L., Winterhalder, B., Cuatianquiz-Lima, C., Rosetti, M. F., Hudson, R., & Ross, C. T. (2019). Nahua mushroom gatherers use area-restricted search strategies that conform to marginal value theorem predictions. *Proceedings of the National Academy of Sciences*, *116*(21), 10339–10347.

Platanios, E. A., Saparov, A., & Mitchell, T. (2020). Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*.

Stephens, D. W., & Krebs, J. R. (2019). *Foraging theory*. Princeton University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tang, W., & Bennett, D. A. (2010). Agent-based modeling of animal movement: A review. *Geography Compass*, *4*(7), 682–700.