Optimal Foraging by Learning the World Model

Roxana Zeraati (research@roxanazeraati.org)
Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Tiffany Oña Jodar (tiffany.ona@alleninstitute.org)
Allen Institute for Neural Dynamics,
Seattle, USA

Shervin Safavi (research@shervinsafavi.org)
Computational Neuroscience, Department of Child and Adolescent Psychiatry,
Faculty of Medicine, TU Dresden, Dresden, Germany
Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Bruno Cruz (bruno.cruz@alleninstitute.org)
Allen Institute for Neural Dynamics,
Seattle, USA

Cindy Poo (cindy.poo@alleninstitute.org)
Allen Institute for Neural Dynamics,
Seattle, USA

Peter Dayan (dayan@tue.mpg.de)
Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics,
Department of Computer Science, University of Tübingen,
Tübingen, Germany

Abstract

Patch foraging—deciding when to leave a depleting resource to search for alternatives—is a fundamental aspect of animal behavior and offers a window into ethologically grounded decision processes. Several theories, most notably the Marginal Value Theorem (MVT), have proposed strategies for optimal foraging. However, they typically ignore most details of the spatiotemporal structure of the environment, and particularly the dynamics of the replenishment of patches. We investigate optimal patch foraging with richer replenishment timescales. Using average-reward reinforcement learning (RL), we show that under slow replenishment, optimal policies leverage the world model to generate higher reward rates and distinct behavioral statistics from MVT and similar policies. Our results provide testable predictions for future experiments.

Keywords: foraging; reinforcement learning; world model

Foraging in naturalistic environments

Foraging is one of the most widespread and evolutionarily conserved behaviors observed across species, making it a powerful window into the cognitive processes underlying natural decision making (Mobbs, Trimmer, Blumstein, & Dayan, 2018). One of the classic forms of foraging is patch foraging, which assumes that food resources are distributed in discrete patches (e.g., fruit bushes) across the environment. Patch foraging can be described as a sequence of decisions as to when to leave a depleting patch to search for better alternatives. The optimal policy for patch foraging is usually described by MVT: an animal should leave a patch when its instantaneous reward rate falls below the average reward rate of the environment (Stephens & Krebs, 1986). While MVT has been successful in describing foraging decisions under certain conditions, it fails to capture all the statistics of foraging behavior (Kendall & Wikenheiser, 2022). In particular, it relies on unrealistic assumptions, e.g., immediate replenishment of a patch to its full capacity after an animal leaves it (or, worse, enters a new patch). Moreover, MVT suggests that only knowing the environment's average reward rate is sufficient to make optimal decisions, discarding the spatiotemporal structure of the environment. Here, we show that in the presence of replenishment with realistic timescales, a complete model of the environment is required to make optimal foraging decisions. Specifically, we employ the average-reward RL framework to find the optimal policy across various environmental statistics and show that it can deviate substantially from MVT. Our results suggest model-based strategies are required for optimal foraging in natural environments.

Environment and task structure

We consider an environment consisting of 3 patches $i \in \{\text{A, B, C}\}$, each with a specific depletion rate D_i, replenishment rate R_i, replenishment timescale τ_i, and maximum and minimum reward probabilities $p_i^{\max}$ and $p_i^{\min}$ respectively. Our RL agent

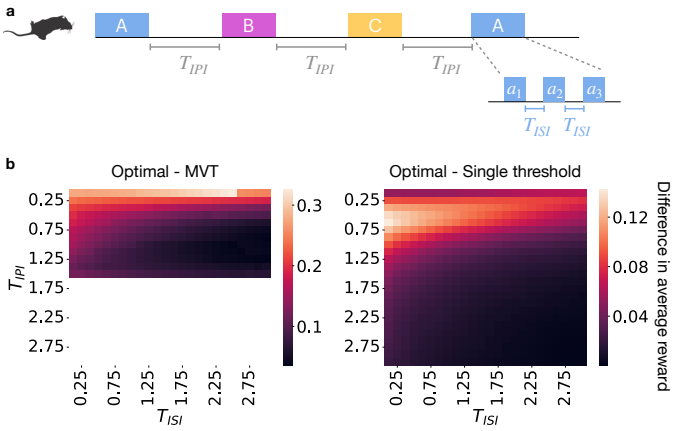

Figure 1: **Impact of environment statistics on foraging policies. a.** Schematic for sequential patch foraging task. Patches provide stochastic rewards with probabilities that decrease when they are exploited, and increase when they are not. **b.** Comparing the average reward rate of the optimal policy with MVT (left) and single-threshold (right) policies. White areas indicate regimes where MVT is impossible. $\tau_i = 3$.

is presented with patches in sequential order, separated by inter-patch-interval T_{IPI} (Figure 1a, analysis can be extended to the case where the patch order is not fixed but has Markov dynamics). Each patch also contains discrete reward sites, separated by inter-site-interval T_{ISI}. For simplicity, we assume T_{IPI} and T_{ISI} are constant. At each time-step t, the agent can decide whether to exploit the current patch or leave. The state of each patch is defined by the reward probability p_i^t. Exploiting a patch can result in a unit reward with probability p_i^t, followed by a handling time t_h for reward consumption, while a leave decision gives zero reward. The reward probability of the prevailing patch is depleted as $p_i^{\max} D_i^n$, where n is the number of received rewards. Meanwhile, the reward probability of non-prevailing patches replenishes as $p_i^{\max} - (p_i^{\max} - p_i^{\text{last}}) R_i^{t/\tau_i}$, where p_i^{last} is the patch state when it was last visited.

Defining policies

First, we find the optimal policy assuming that the agent has complete knowledge of the environment statistics. We formulate the foraging problem as a Markov Decision Process (MDP). The states are defined as $s(l^t, p_{\text{A}}^t, p_{\text{B}}^t, p_{\text{C}}^t)$, where l^t is the prevailing patch. The state transitions $p(s'|s,a)$ are defined based on the agent's actions $a \in \{\text{exploit, leave}\}$, and patch depletion and replenishment dynamics. We find the optimal policy by using policy iteration to solve the average-reward Bellman equation (Mahadevan, 1996),

$$V^*(s) = \max_a \{ r(s,a) - T(s,a)\bar{r}^* + \sum_{s'} p(s'|s,a) V^*(s') \}.$$

Here, V^* is the optimal state-value function, r^* is the optimal average reward, and $T(s,a)$ is the time it takes to perform action a at state s: $T(s, \text{exploit}) = T_{\text{ISI}}$ and $T(s, \text{leave}) = T_{\text{IPI}}$. For rewarded exploitation, t_h is added to $T(s, a = \text{exploit})$. All

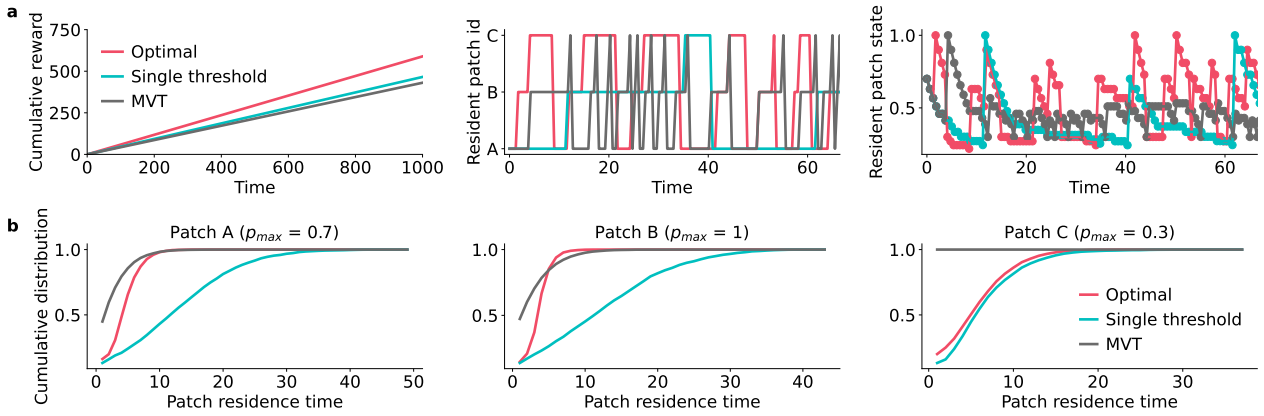

Figure 2: **Policies have different behavioral predictions.** Different policies obtain (**a**) distinct amounts of reward, visit distinct patches and patch states with (**b**) different patch residence times (all simulations start from fully replenished environment i).

times are defined in relative terms; we assume $t_h = 0.25T_{ISI}$. $T(s,a)\bar{r}^*$ captures the opportunity cost of time during foraging.

We also define two suboptimal policies: (i) MVT: a self-consistent policy with a leaving threshold equal to the average reward; (ii) Optimal single threshold: a policy that maximizes the average reward rate by applying a fixed leave threshold across all patches.

Comparing different policies

We compare the three policies by their average reward rate, visited patches and patch states, and the distribution of patch residence times, across different environmental statistics. In particular, we search for environment statistics that maximize the difference in the average reward rate between policies. As examples, here, we present the results in two types of environments: (i) patches with heterogeneous maximum reward probabilities $p_i^{\max} = \{0.7, 1, 0.3\}$ and (ii) uniform patch statistics $p_i^{\max} = 0.7$. We set $R_i = D_i = 0.9$, $p_i^{\min} \geq 0.2$ and $\tau_i = 3$ and perform a grid search on T_{IPI} and T_{ISI} to find regimes with maximum difference between policies.

We find that when the replenishment timescale is much slower than the inter-patch interval (and to a lesser extent also inter-reward-site interval), the optimal policy substantially deviates from MVT and single-threshold policies (Figure 1b for environment i; qualitatively similar results for environment ii). Under this condition, patches do not fully replenish until subsequent visits, violating the MVT assumption. Thus, the agent requires a richer model of the environment beyond the average reward rate to make optimal decisions, even when in a uniform environment. Moreover, we find that single-threshold and MVT policies deviate from each other, indicating that the environment's average-reward rate does not provide sufficient information even for choosing an optimal single leave threshold. The difference between the optimal and the other two policies lies in its adaptive leave thresholds. Unlike MVT and single-threshold policies that apply a fixed leave threshold across all patches, the optimal policy adapts its leave threshold according to the state of all 3 patches at each moment of time. We also show that in certain environments, it would not be possible to define a self-consistent MVT policy due to a small overall average reward rate (and large $p_i^{\min}$, white areas in Figure 1b). Our results can generally be extended to other parameter sets as long as T_{IPI} is considerably smaller than τ_i.

Differences between policies provide testable behavioral predictions. To illustrate this, we take an example environment where policies have different average reward rates (environment i with $\tau_i = 3$, $T_{\text{IPI}} = 0.625$, $T_{\text{ISI}} = 0.5$). We find that policies not only obtain different amounts of reward, but also consistently visit different patches and patch states (Figure 2a). Furthermore, they exhibit different distributions of patch residence times (time spent exploiting a patch). In our specific example environment, the MVT policy almost skips patch C, while the other two policies spend more time there (Figure 2b). Such a striking difference provides a strong testable prediction to distinguish between MVT and optimal policies in experiments.

Our results demonstrate that under slow replenishment timescales, foraging strategies informed by the world model provide a higher reward rate and are optimal for survival. Since slow replenishment is a realistic property of natural environments, we hypothesize that animals learn the world model and leverage it to inform their foraging decisions (e.g., using model-based RL). In the next step, we plan to test this hypothesis by testing mice in a naturalistic patch-foraging paradigm designed based on our theoretical predictions.

Acknowledgments

This work was supported by the Max Planck Society (R.Z., S.S., P.D.), the add-on fellowship from the Joachim Herz Foundation (R.Z., S.S.), the Allen Institute for Neural Dynamics (T.O.J., B.C., C.P.), the German Research Foundation (DFG) grant 550411021 (S.S.), and the Humboldt Foundation (P.D.).

References

Kendall, R. K., & Wikenheiser, A. M. (2022). Quitting while

you're ahead: Patch foraging and temporal cognition. *Behavioral Neuroscience*, *136*(5), 467.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, *22*(1), 159–195.

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: insights from ethology. *Nature Reviews Neuroscience*, *19*(7), 419–427.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton university press.