A way around the exploration-exploitation dilemma.

Erik J Peterson (erik.exists@gmail.com)
Dept. of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15213

Timothy Verstynen (timothyv@andrew.cmu.edu)
Dept. of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15213

Abstract

The exploration-exploitation dilemma is considered a fundamental but intractable problem in the learning and decision sciences. At its crux is the search to maximize reward. Here we challenge this view and show a way around the dilemma by defining separate mathematical objectives for exploration and exploitation. To make the objective for exploration independent of reward, we derive a set of general axioms for information value. Using these axioms we develop a greedy algorithm which provably and optimally maximizes both information and reward.

Keywords: reinforcement learning; exploration; theory

Introduction

Exploration during behavior can have two very different explanations depending on whether and animal or aritifical agent might recieve a reward. If there is no reason to expect a reward, exploration is treated as a search for novel or maximum information (). For example, when a rat is placed in a new environment it will explore even if no tangible rewards, like food and water, are present or expected (Liu et al., 2019; ?, ?, ?). If however reward is expected an animal, and many artificial agents, will explore and discover it (?, ?). This exploration gets interpreted as a search to maximize reward (Sutton & Barto, 2018).

An open problem for the decision sciences is to provide a general and unified account for both kinds of exploration. This is difficult in part because exploration for reward leads to the intractable problem of the exploration-exploitation dilemma, which is fundamental (Thrun, 1992; Dayan & Sejnowski, 1996; Findling, Skvortsova, Dromnelle, Palminteri, & Wyart, 2018; Gershman, 2018).

But is it really fundamental?

Results

We conjecture *all* exploration can be correctly interpreted as a search for information. Reward value doesn't matter. Using this conjecture we'll show there exists a tractable solution to maximize reward value which is highly general, provably optimal, and strictly deterministic.

If exploration is *just* a search for max information, then what was the "dilemma" can be divided into two problems. One problem is exploration (recast as the search for max information) and the other is exploitation, which still means a search for max reward. The remainder of this paper works toward deriving an simple greedy algorithm for solving both problems simultaneously.

A definition of information value

To make our eventual solution general, we first need to separate reward value from information value in a principled and general way. We do this axiomatically. Our axioms *do not* depend on information theory nor on Bayesian reasoning, the standard approaches (Friston et al., 2016; Haarnoja et al., 2018; Itti & Baldi, 2009).

We reason that the value of any observation s made an animal or agent dependents entirely on what an animal or agent learns by making that observation–how it changes an animal or agent's memory.

We let a memory M be finite set whose maximum size is N–defined over a finite state space $s \in S^N$–whose elements are added by an invertible encoder f, such that $M_{t+1} = f(M_t, s)$ and $M_t = f^{-1}(M_{t+1}, s)$ and whose elements z from M are recovered by a decoder g, such that $z = g(M, s)$. The initial memory M_0 is said to be the empty set, $M_0 = \emptyset$ (Definition 1).

Our notion of memory is general. We don't say anything about the details of the encoder f, the decoder g, the numerical properties of the individual states s in the state-space S (other than S is finite). Nor do we require that what goes into M (i.e., s) and what comes out (z) have any fixed relationship.

Axiomatic information value (E)

Axiom 1 (Axiom of Now)**.** *E is depends only on difference dM between M_{t+1} and M_t.*

That is, the value of an observation s depends only on how the current memory changes. Its history doesn't matter.

Axiom 2 (Axiom of Novelty)**.** *$E = 0$ if and only if $dM = 0$.*

That is, an observation that doesn't change the memory has no value.

Axiom 3 (Axiom of Scholarship)**.** *$E \geq 0$.*

That is, all (new) information is *in principle* valuable even it's consequences are later found to be negative.

Axiom 4 (Axiom of Specificity)**.** *E is monotonic with the compactness C of dM (Eq. 6) .*

That is, more specific information is more valuable then less specific information. (We use the mathematical idea of compactness to formalize specificity.)

Axiom 5 (Axiom of Equilibrium). $\frac{dM^2}{d\theta^s} < 0$

That is, using the parameters θ learning in M makes continual progress toward equilibrium (i.e. self-consistency) with each observation and will approach a steady-state value of $dM \to 0$ and $E \to 0$.

Though they are formally independent, the axioms do borrow properties from Bayesian learning and information geometry (Schwartenbeck et al., 2019; Friston et al., 2016; Harper, 2009; Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017; Itti & Baldi, 2009). As a result, our def. of axiomatic value reduces to the KL divergence/information gain (MacKay, 2003; Ly et al., 2017) if one can assume the animals or agents memory M is Bayesian or otherwise a probability distribution.

Exploration as a dynamic programming problem

A nice solution to maximizing E would be a dynamic programming solution. It would guarantee that total value (Eq. 2) is maximized by a simple deterministic, locally greedy, algorithm.

In Theorem 1 we prove our definition of memory (Def. 1) has one critical property (optimal substructure) needed for a dynamic programming solution (?, ?). The other two–$E \geq 0$ and the Markov property(Sutton & Barto, 2018)–are fulfilled by the Axioms 3 and 1 respectively.

To write down the Bellman solution for E as dynamic programming problem we need some new notation.

Let a be an action drawn from a finite action space A^K. Let time t be denote an index $t = (1, 2, 3, \ldots, T-1)$. Let π be a policy function that maps a state s to an action a, $\pi : s \to a$, such that $\forall s_t \in S, \forall a_t \in A$. Let δ be a transition function which maps (s_t, a_t) to a new state s_{t+1}, $\delta : (s_t, a_t) \to s_{t+1}$.

For notational simplicity we let E be redefined as $F(M_t, a_t)$–a "payoff function" (Eq 1).

$$\begin{aligned} F(M_t, a_t) &= E(M_{t+1}, M_t) \\ &\text{subject to the constraints} \\ a_t &= \pi(s_t) \\ s_{t+1} &= \delta(s_t, a_t), \\ M_{t+1} &= f(M_t, s_t) \end{aligned} \tag{1}$$

The value function for E is Eq 2.

$$V^*_{\pi_E}(M_0) = \max_{a \in A} \sum_{t \in T} F(M_t, a_t) \tag{2}$$

To find recursive Bellman solution we decompose Eq. 2 into an initial value F_0 and the remaining series. When this decomposition is applied recursively we eventually find an iterative optimal and greedy solution. The steps for our decomposition in terms of F and M are shown in Eq 4 and the result in Eq. 3.

$$\begin{aligned} V^*_{\pi_E}(M_0) &= \max_{a \in A} \Big[\sum_{t \in T} F(M_t, a_t) \Big] \\ &= \max_{a \in A} \Big[F(M_0, a_0) + \sum_{t \in T} F(M_{t+1}, a_{t+1}) \Big] \\ &= F(M_0, a_0) + \max_{a \in A} \Big[\sum_{t \in T} F(M_{t+1}, a_{t+1}) \Big] \\ &= F(M_0, a_0) + V^*_{\pi_E}(M_{t+1}) + V^*_{\pi_E}(M_{t+2}), \ \ldots \end{aligned} \tag{3}$$

$$V^*_{\pi_E}(M_t) = F(M_t, a_t) + \max_{a \in A} \Big[[F(M_{t+1}, a_t) \Big] \tag{4}$$

A note on exploration quality

Having a policy π^*_E that maximizes E also does not necessarily ensure that exploration is of good quality. Relying on Axiom 6, Theorems 2 and 3 prove π^*_E also leads to a complete and exhaustive exploration of any *finite* space S.

- (Theorem 2.) The optimal policy π^*_E must visit each state in $s \in S$ at least once.
- (Theorem 3.) The optimal policy π^*_E must revisit each $s \in S$ until learning about each state s until the memory reaches equilibrium $dM = 0$.

Scheduling a way around the dilemma

To solve the dual problems we've posed–information and reward maximization–we need an algorithm the can maximize the total value of both objectives. We found a simple (if surprising solution) in the computer science sub-field of optimal scheduling.

There are range of optimal and useful reinforcement learning algorithms (Sutton & Barto, 2018). In principle, any of them is compatible with our approach. As a result we use π_R to denote any reinforcement learning algorithm that operates on a discrete Markov decision space (Sutton & Barto, 2018).

Animal or agent behavior can be generally viewed as a fixed resource in at it can only take one action at a time. With this in mind, we imagine the policies for exploration and exploitation act as two possible "jobs" competing for control of this fixed (behavioral) resource. By definition each of these "jobs" naturally produces non-negative values a optimal job scheduler could use. E for information or R reward/reinforcement learning. Each of the "jobs" takes a constant amount of time, each policy only results in a single action $a \sim A$.

The solution to scheduling problems that 1. produce non-negative value and 2. have fixed run times is known to be a simple greedy algorithm (Roughgarden, 2019). We restate this solution as a pair of inequalities in Eq. 5.

$$\begin{aligned} \pi_\pi &= \begin{cases} \pi^*_E & : E_t - \varepsilon > R_t \\ \pi_R & : E_t - \varepsilon \leq R_t \end{cases} \\ &\text{subject to the constraints} \\ R &\in \{0, 1\} \\ p(R) &< 1 \end{aligned} \tag{5}$$

The fine print The inequality in Eq 5 ensures total value summed over E_t and R_t is maximized (Roughgarden, 2019). It does not ensure each policy also maintains its own optimality. Specifically we're interested in Theorems 2 and 3. We ensure these hold by introducing two constraints the reward distribution shown in Eq 5 (and which we prove are sufficient in Theorem 4).

In stochastic environments learning in M may show small continual fluctuations, mirroring the environmental dynamics. These will of course get reflected in E fluctuations. To allow Eq. 5 to achieve a stable solution we introduce ε, a boredom threshold for exploration. When $E_t < \varepsilon$ we say the policy is "bored" with exploration. To be consistent with the value axioms, $\varepsilon > 0$ and $E + \varepsilon \geq 0$.

To ensure the default policy is reward maximization, Eq. 5 breaks ties between R_t and E_t in favor of π_R.

Limitations

The initial value E_0 for π^*_E can be arbitrary with the limit $E_0 > 0$. In theory E_0 does not change π^*_E's long term behavior, but different values will change the algorithms short-term dynamics and so might be quite important in practice.

By definition a pure greedy policy, like π^*_E, can't handle ties. There is simply no mathematical way to rank equal values. Theorems 3 and 2 ensure that any tie breaking strategy is valid. However like the choice of E_0, tie breaking can strongly effect the transient dynamics.

Algorithmic complexity

The worst case run time for π_π is additive in its policies. That is, if in isolation it takes T_E steps to earn $E_T = \sum_{T_E} E$, and T_R steps to earn $r_T = \sum_{T_R} R$, then the worst case training time for π_π is $T_E + T_R$, if that is neither policy can learn from the other's control. There is however no reason each policy can't observe the transitions (s_t, a_t, R, s_{t+1}) caused by the other. If this is allowed, worst case training time improves to $\max(T_E, T_R)$.

Summary

We show a way around the exploration-exploitation dilemma. In our approach exploration is done only to maximize information value, a quantity we define axiomatically. Maximizing information value forces the animal to learn a general, reward-independent, memory of the world. An important "side effect" of this general learning is that reward learning improves to optimality.

References

Dayan, P., & Sejnowski, T. J. (1996, October). Exploration bonuses and dual control. *Machine Learning*, *25*(1), 5-22. doi: 10.1007/BF00115298

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2018, October). *Computational noise in reward-guided learning drives behavioral variability in volatile environments* (Preprint). Neuroscience. doi: 10.1101/439885

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., ODoherty, J., & Pezzulo, G. (2016, September). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862-879. doi: 10.1016/j.neubiorev.2016.06.022

Gershman, S. J. (2018, April). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34-42. doi: 10.1016/j.cognition.2017.12.014

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., ... Levine, S. (2018, December). Soft Actor-Critic Algorithms and Applications. *arXiv:1812.05905 [cs, stat]*.

Harper, M. (2009, November). The Replicator Equation as an Inference Dynamic. *arXiv:0911.1763 [cs, math]*.

Itti, L., & Baldi, P. (2009, June). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295-1306. doi: 10.1016/j.visres.2008.09.007

Liu, A., Papale, A., Hengenius, J., Patel, K., Ermentrout, B., & Urban, N. (2019, February). Mouse navigation strategies for odor source localization. *bioRxiv*. doi: 10.1101/558643

Ly, A., Marsman, M., Verhagen, J., Grasman, R., & Wagenmakers, E.-J. (2017, May). A Tutorial on Fisher Information. *arXiv:1705.01064 [math, stat]*.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms* (2nd ed.).

Roughgarden, T. (2019). *Algorithms Illuminated (Part 3): Greedy Algorithms and Dynamic Programming* (Vol. 1) (No. 3).

Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *eLife*(e41703), 45.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition ed.). Cambridge, Massachusetts: The MIT Press.

Thrun, S. B. (1992). Eficient Exploration In Reinforcement Learning. *NIPS*, 44.

Mathematical Appendix.

Compactness

The compactness C of a hyper-cube has a simple formula, $C = \frac{P^2}{A}$ where P is the cube's perimeter and A is its area. Therefore as M_t moves to M_{t+1}, the we can measure the distances $\{d_i, d_{i+1}, d_{i+2}, \dots d_N\}$ for all $O \leq N$ observed states Z, such that $Z \subseteq S$ and treat them as if they formed a O-dimensional hyper-cube. In measuring this imagined cube we arrive at a geometric estimate for change the compactness of our memory M (Eq. 6).

$$C = \frac{\left(2\sum_O^i d_i\right)^2}{\prod_O^i d_i} \tag{6}$$

Information value as a dynamic programming problem

To use theorems from dynamic programming (Roughgarden, 2019; Sutton & Barto, 2018) we must prove our memory M has optimal substructure. By optimal substructure we mean that M can be partitioned into to a small number collection or series, each of which is an optimal dynamic programming solution. In general by proving we can decompose some optimization problem in a set of smaller problems whose optimal solution is easy to find or prove, it trivial to prove we can also grow the series optimally which in turns allows for direct proofs by induction.

Theorem 1 (Optimal substructure)**.** *Assuming transition function δ is deterministic, if $V^*_{\pi_E}$ is the optimal information value given by π_E, a memory M_{t+1} has optimal substructure if the the last observation s_t can be removed from M_t, by $M_{t+1} = f^{-1}(M_{t+1}, s_t$ where the resulting value $V^*_{t-1} = V^*_t - F(M_t, a_t)$ is also optimal.*

Proof. Given a known optimal value V^* given by π_E we assume for the sake of contradiction there also exists an alternative policy $\hat{\pi}_E \neq \pi_E$ that gives a memory $\hat{M}_{t-1} \neq M_{t-1}$ and for which $\hat{V}^*_{t-1} > V^*_{t-1}$.

To recover the known optimal memory M_t we lift $\hat{M}_{t-1}$ to $M_t = f(\hat{M}_{t-1}, s_t)$. This implies $\hat{V}^* > V^*$ which in turn contradicts the purported original optimality of V^* and therefore $\hat{\pi}_E$. □

A greedy policy explores exhaustively

Our proofs for exploration breadth are really sorting problems. If every state must be visited (or revisited) until learned, then under a greedy policy every state's value must–at one time or another–be the maximum value.

Let Z be set of all visited states, where Z_0 is the empty set $\{\}$ and Z is built iteratively over a path P, such that $Z = \{s | s \in P$ and $s \notin Z\}$.

Sorting requires ranking, so we define an algebraic notion of inequality for any three numbers $a, b, c \in \mathbb{R}$ are defined in Eq. 7.

$$a \leq b \Leftrightarrow \exists c; b = a + c \tag{7}$$

$$a > b \leftrightarrow (a \neq b) \wedge (b \leq a) \tag{8}$$

Theorem 2 (State search – completeness and uniqueness)**.** *A greedy policy π is the only deterministic policy which ensures all states in S are visited, such that $Z = S$.*

Proof. Let $\mathbf{E} = (E_1, E_2, ...)$ be ranked series of E values for all states S, such that $(E_1 \geq E_2, \geq ...)$. To swap any pair of values $(E_i \geq E_j)$ so $(E_i \leq E_j)$ by Eq. 7 $E_i - c = E_j$.

Therefore, again by Eq. 7, $\exists \int \delta E(s) \rightarrow -c$.

Recall: $\nabla \mathcal{L}_M < 0$

However if we wished to instead swap $(E_i \leq E_j)$ so $(E_i \geq E_j)$ by definition $\nexists c; E_i + c = E_j$, as $\nexists \int \delta \rightarrow c$.

To complete the proof, assume that some policy $\hat{\pi}_E \neq \pi^*_E$. By definition policy $\hat{\pi}_E$ can any action but the maximum leaving $k-1$ options. Eventually as $t \rightarrow T$ the only possible swap is between the max option and the kth but as we have already proven this is impossible as long as $\nabla \mathcal{L}_M < 0$. Therefore, the policy $\hat{\pi}_E$ will leave at least 1 option unexplored and $S \neq Z$. □

Theorem 3 (State search – convergence)**.** *Assuming a deterministic transition function Λ, a greedy policy π_E will resample S to convergence as $t \rightarrow T$, $E_t \rightarrow 0$.*

Proof. Recall: $\nabla \mathcal{L}_M < 0$.

Each time π^*_E visits a state s, so $M \rightarrow M'$, $F(M', a_{t+1}) < F(M, a_t)$

In Theorem 2 we proved only deterministic greedy policy will visit each state in S over T trials.

By induction, if $\pi^* E$ will visit all $s \in S$ in T trials, it will revisit them in $2T$, therefore as $T \rightarrow \infty$, $E \rightarrow 0$. □

Optimality of π_π

Theorem 4 (Optimality of π_π)**.** *Assuming an infinite time horizon, if π_E is optimal and π_R is optimal, then π_π is also optimal in the same sense as π_E and π_R.*

Proof. The optimality of $|\pi_\pi$ can be seen by direct inspection. If, $p(R = 1) < 1$ and we have an infinite horizon, the π_E will have a unbounded number of trials meaning the optimally of P^* holds. Likewise, $\sum E < \varepsilon$ as $T \rightarrow \infty$, ensuring pi_R will dominate π_π therefore π_R will asymptotically converge to optimal behavior. □

In proving the total optimality of π_π we limit the probability of a positive reward to less than one, denoted by $p(R_t = 1) < 1$. Without this constraint the reward policy π_R would always dominate π_π when rewards are certain. While this might be useful in some circumstances, from the point of view π_E it is extremely suboptimal as the model would never explore. Limiting $p(R_t = 1) < 1$ is reasonable constraint, as rewards in the real world are rarely certain. A more naturalistic but complex way to handle this edge case might be to introduce reward satiety, and have reward value decay with repeated exposure.