
Active Reinforcement Learning: Observing Rewards at a Cost

David Krueger
Montreal Institute for Learning Algorithms
University of Montreal
david.krueger@umontreal.ca

Jan Leike
Future of Humanity Institute
University of Oxford
jan.leike@philosophy.ox.ac.uk

Owain Evans
Future of Humanity Institute
University of Oxford
owain.evans@philosophy.ox.ac.uk

John Salvatier
AI Impacts
jsalvatier@gmail.com

Abstract

Active reinforcement learning (ARL) is a variant on reinforcement learning where the agent does not observe the reward unless it chooses to pay a *query cost* $c > 0$. The central question of ARL is how to quantify the long-term value of reward information. Even in multi-armed bandits, computing the value of this information is intractable and we have to rely on heuristics. We propose and evaluate several heuristic approaches for ARL in multi-armed bandits and (tabular) Markov decision processes, and discuss and illustrate some challenging aspects of the ARL problem.

1 Introduction

Translating human objectives into rewards for a reinforcement learning (RL) agent is often challenging and time-consuming. Tasks such as driving in a busy city or engaging in a conversation have complex reward structures which humans understand via high-level concepts such as “safety” and “enjoyment”. For such tasks, it may be prohibitively expensive to design a reward function by hand for an agent that does not already understand such concepts.

Instead of specifying a reward function in advance, a human can evaluate state-actions as they occur and provide reward to the agent *online*. In this case, it’s not necessary to assign rewards to all possible state-actions, only to those which are visited. Manually providing these rewards may still be labor intensive and costly, however. For instance, in a clinical trial the rewards for an RL agent are patient outcomes, measured by expensive medical tests. Likewise, it may be expensive to obtain high quality feedback from users of a web application.

To account for the cost of collecting reward feedback, we consider a simple modification to traditional RL called *active reinforcement learning* (ARL). In analogy with active learning, an ARL agent *chooses* when to observe the reward signal.

Informal problem statement for ARL:

At each time-step, the agent chooses both an action and whether to observe the reward in the next time-step. If the agent chooses to observe the reward, then it pays the “query cost” $c > 0$. The agent’s objective is to maximize total reward minus total query cost.

By intelligently choosing which rewards to observe, ARL agents can learn more efficiently about the reward function. As in active learning, the agent can exploit statistical dependencies between the

rewards of different state-actions, observing rewards that are expected to be most informative about other rewards. Unlike in active learning, the goal in ARL is not to learn the reward function, but rather to take actions that maximize the discounted sum of rewards. Towards this end, ARL agents can also exploit knowledge of the environment’s dynamics to query the state-actions most relevant to improving their policy.

The ARL problem involves two important simplifying assumptions:

1. Rewards are observed immediately (i.e. without any delay)
2. The agent only ever observes rewards for the *current* state-action (not past state-actions)

Relaxing these assumptions would be valuable but is beyond the scope of this paper.

For many problems, requiring a human to be “on-call”, i.e. ready to provide rewards without delay, is itself costly. This could be addressed by using a query cost which is situation dependent, e.g. determined online by a human. However, when the agent acts on time-scales much shorter than the human, it’s not possible to provide it instantaneous rewards. Further work could explore the consequences of reward signals being delayed.

Restricting queries to the current state-action is natural if the agent’s representation of state is missing some important features of the environment which the human can observe, for example in partially observable problems. In this setting, the human can make use of their additional information when providing rewards. If the agent’s state-representation captures all important features, then it could query the human about past or hypothetical state-actions. Future work will explore this case.

This paper presents preliminary work on active reinforcement learning (ARL). We propose and evaluate several heuristic algorithms for ARL in multi-armed bandits and tabular MDPs. We present examples, theoretical observations and empirical comparisons that shed light on the character of the ARL problem. The central question of ARL is how to quantify the long-term value of reward information. As the cost of observing rewards gets high (relative to possible discounted future rewards) the problem may become quite different from the traditional RL problem with fully-observed reward.

1.1 Related work

The difficulty of encoding human objectives in reward functions has been articulated and addressed in a substantial literature. This includes work in inverse reinforcement learning (Ng and Russell, 2000; Evans et al., 2015) and preference-based reinforcement learning (Wirth and Fürnkranz, 2013).

The TAMER framework (Knox and Stone, 2009) has a similar motivation to ARL. The authors consider the setting where rewards are provided online by a human “teacher” and explicitly take into account the time delays and noise in the human’s responses. TAMER typically assumes that the human provides signals about the *value* of actions (e.g. $Q^*(s, a)$), rather than their rewards (as in ARL).

Recent work on “Active Reward Learning” (Daniel et al., 2014) also shares its motivation with ARL. Daniel et al. test RL with online human feedback empirically, with humans evaluating how well a robot grasps objects. Their algorithm is designed for continuous control tasks, where the reward function is defined on entire trajectories rather than individual actions. They employ an active learning approach based on Gaussian Process regression and Bayesian Optimization. However, their techniques are not readily applicable to the discrete problems we study here.

Finally, there is a wide range of work in which an RL agent performs some kind of active learning of reward information in a setting that is less similar to ARL (Weng et al., 2013; Lopes and Montesano, 2014; Regan and Boutilier, 2011).

2 Active Bandits

This section formally introduces and discusses the active reinforcement learning problem in the context of multi-armed bandits.

2.1 Problem Formulation

The *active multi-armed bandit problem* is an online learning problem where the learner has to select from K actions (‘arms’) in every round. Choosing action $A_t \in \{1, \dots, K\}$ in round t yields a reward R_t of drawn from a time-independent distribution ν_{I_t} with mean μ_i unknown to the learner. In contrast to the classical multi-armed bandit problem, the learner does not observe the reward unless they pay a fixed *query cost* $c > 0$. This cost is the same for every arm and every round.

Let $\mu^* := \max_i \mu_i$ denote the mean of the best arm. The goal of the learner is to maximize expected reward up to a known horizon n , or, equivalently, to minimize the *expected regret*

$$\text{Regret}_n = \mathbb{E} \left[\sum_{t=1}^n (\mu^* - R_t - cQ_t) \right],$$

where Q_t is 1 if the learner chooses to observe the reward in round t and 0 otherwise. In other words, the expected regret is the expected reward that is lost because the learner did not blindly choose the best option in every round.

2.2 Properties

The active multi-armed bandit problem is a subproblem of the partial monitoring problem (Piccolboni and Schindelhauer, 2001): the action space consists of the $2K$ actions (the K arms with and without querying) and the feedback is the observed payoff or a blank symbol. In contrast to most of the literature on partial monitoring, we consider the environment to be stochastic instead of adversarial. For distributions with finitely many possible rewards (such as Bernoulli distributions), active multi-armed bandits are stochastic finite partial monitoring problems (Komiyama et al., 2015). In case of two outcomes, (adversarial) partial monitoring problems can be classified in four different categories: 0, $\Theta(n^{1/2})$, $\Theta(n^{2/3})$, or $\Theta(n)$ regret (Antos et al., 2013). These categories are called trivial, easy, hard, and hopeless respectively. While multi-armed bandits fall into the easy category, *active* multi-armed bandits fall into the hard category: their worse-case expected regret is $\inf_{\pi} \sup_{\mu_1, \dots, \mu_K} \text{Regret}_n \in \Theta(n^{2/3})$.

Intuitively, the fundamental difference to (regular) multi-armed bandit problems is in the cost of distinguishing two very close arms. In active bandits, distinguishing two close arms incurs regret linear in the number of time steps needed to distinguish them ($n^{2/3}$ in the worst case), whereas in (regular) bandits the regret grows proportionally to the gap size ($n^{1/2}$ in the worst case).

To achieve optimal asymptotic regret, we can ignore the magnitude of the query cost, since it is constant. Partial monitoring algorithms are designed for optimal asymptotic regret rate and do not take the magnitude of the query cost into account. Here we are particularly interested in optimizing the cost-dependent regret.

The optimal strategy is to query for a number of time steps and then stop querying altogether. If you don’t query, then in the next step you face the same information, except that the horizon is now shorter, so the value of information can only have diminished.

Active multi-armed bandit problems degenerate into two familiar problems in the edge cases. If $c = 0$ we face a regular bandit problem and the regret Regret_n corresponds to the cumulative regret. If $c \gg \max_i \Delta_i$ where $\Delta_i := \max_j \mu_j - \mu_i$ is the gap size, then Regret_n is dominated by the simple regret (the gap of the chosen arm), and so we have a best arm identification problem¹. But there is a trade-off between simple regret and cumulative regret (Bubeck et al., 2011, Thm. 1): if the cumulative regret is small, then there is a lower bound on the simple regret. Any algorithm for a moderate cost setting therefore has to manage the balancing act between cumulative and simple regret.

2.3 Algorithms

For finite sets of rewards, PM-DMED achieves the optimal cost-independent regret rate (Komiyama et al., 2015). We are looking for a (heuristic) algorithm that

- achieves optimal asymptotic (cost-dependent) regret rate,

¹For $c \gg 0$, it may be optimal to never query and simply chose the best arm according to the prior

- performs well in practice, and
- has constant or linear time complexity.

Ideally, the algorithm’s performance does not degrade at the edge cases with $c = 0$ and $c \gg \max_i \Delta_i$.

A natural choice is an explore-then-exploit algorithm, like for *budgeted bandits* (Madani et al., 2004). The central question is how to decide to stop querying. In other words, what is the (long-term) value of information of an additional query?

If ν_1, \dots, ν_K are Bernoulli distributions, then the Bayes-optimal solution can be found with dynamic programming: The corresponding belief MDP has $\mathcal{O}(n^{2K})$ states and can be solved in $\mathcal{O}(n^{2K})$ time steps. This is polynomial for constant K , but completely prohibitive in practice.

It is well-known that to solve partial monitoring problems it is sometimes necessary to take actions that you strongly believe to be suboptimal to get information. For active bandits, paying to see the reward is always suboptimal by at least $c > 0$. Nevertheless, if we never pay the cost, we learn nothing about the problem. Because of this, strategies like optimism or Thompson sampling cannot tell us when to query. Moreover, there can be no index strategy² because the decision between two arms can depend on the value of a third arm (Hay et al., 2012, Ex. 4). Lastly, active multi-armed bandits do not allow greedy solutions: Usually one additional data point does not change our opinion which arm we currently consider to be best. A greedy strategy considers the value of information to be too low (or even zero) and stops querying prematurely (Powell and Ryzhov, 2012, Sec. 5.2). Therefore we have to take into account the long-range effects of querying for multiple time steps.

If ν_1, \dots, ν_K are Gaussian distributions, then we can estimate the value of paying the query cost for multiple time steps using a knowledge gradient (Powell and Ryzhov, 2012, Ch. 5). As far as we know, how to efficiently compute the multi-step knowledge gradient for Bernoulli bandits is still an open question.

We introduce a new algorithm called *mind-changing cost heuristic* (MCCH). Let \hat{m} be an estimate of the expected number of time steps we need to query in succession to move the posterior mean of the second best arm to the posterior mean of the best arm (we use $\hat{m} := \max\{1, \lceil 2 \min_i ((T_i + 1) \hat{\Delta}_i)^2 \rceil\}$). Let $\text{Regret}_n(i)$ denote the Bayes-expected regret when committing to arm i now (never querying again) and let \hat{i} denote the current estimate of the best arm. The algorithm pays the query cost if and only if

$$c\hat{m} < \alpha \text{Regret}_n(\hat{i}) \tag{1}$$

where $\alpha > 0$ is a hyperparameter. If (1) holds, then the action is selected by some standard bandit algorithm, such as DMED (Honda and Takemura, 2010). If (1) does not hold, then the algorithm commits to the current best arm for the remaining time steps without paying the query cost.

2.4 Experiments

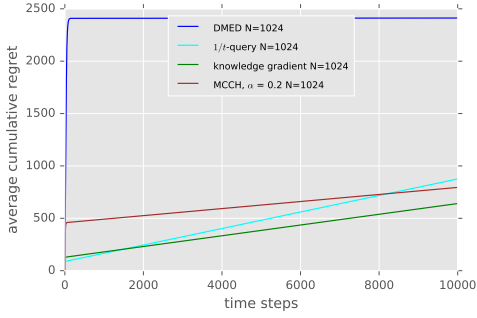
Figure 1 shows the average cumulative regret of different active bandit algorithms for high ($c = 50$) and moderate ($c = 2$) query costs. All of these algorithms use DMED (Honda and Takemura, 2010) to select actions. To be able compute the non-greedy knowledge gradient policy, we approximate prior, posterior, and likelihood with Gaussian distributions. While MCCH never quite beats the other algorithms, it performs more robustly for different query costs. Figure 2 shows that the choice of its hyperparameter α is not very sensitive, which makes it better suited than an optimized fixed (problem-dependent) query stopping time.

3 Active RL in MDPs

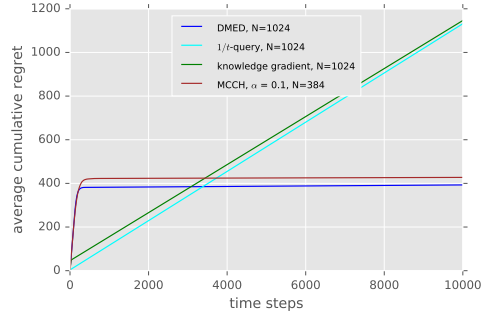
This section studies Active Reinforcement Learning in tabular MDPs³. We begin by reviewing the problem. At every timestep t , the agent chooses an action $A_t \in \mathcal{A}$ and chooses whether to observe, or *query*, the reward sample R_t . This decision is made after choosing the action but before seeing the next state. Choosing to query incurs a cost, which we assume to be a constant $c > 0$ that is known to

²An index strategy computes for each arm a number that is independent of the other arms and takes the action with highest number.

³We use the notational standard MDPNv1 (Philip S. Thomas, 2015)



(a) Means 0.8 and 0.5, cost $c = 50$.



(b) Means 0.7, 0.5, 0.4, 0.4, 0.4, 0.4, cost $c = 2$.

Figure 1: Average cumulative regret for different active bandit algorithms for Bernoulli arms with horizon 10^4 . We compare our algorithm MCCH with knowledge gradient (Powell and Ryzhov, 2012, Ch. 5), querying with probability $1/t$, and a DMED (Honda and Takemura, 2010) variant that stops querying when the algorithm selects only one arm. The latter two algorithms do not take the query cost into account and this is why they sometimes perform poorly.

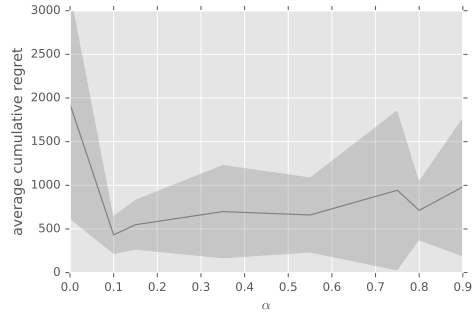
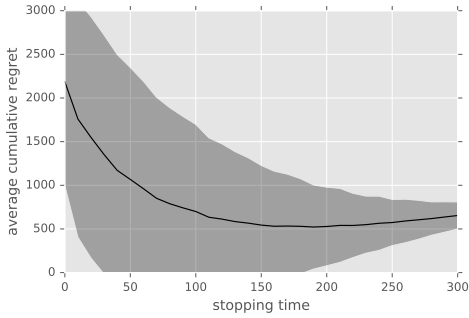


Figure 2: Active 6-armed Bernoulli bandit with means 0.7, 0.5, 0.4, 0.4, 0.4, 0.4, horizon $n = 10^4$, and cost $c = 2$. We compare two policies: DMED with a prespecified query stopping time (left) and MCCH for different values of the hyperparameter α (right). The shaded area corresponds to one standard deviation. MCCH achieves slightly better mean regret, but also is much more robust to the choice of the hyperparameter.

the agent. An ARL agent seeks to maximize the (discounted) sum of rewards (including unobserved rewards), minus the (discounted) sum of the query costs.

An optimal agent for ARL would query a state-action if the Expected Value of Sample Information (EVSI) (Raiffa and Schlaifer, 1961) from the reward observation is larger than the query cost c . This is similar to the standard RL problem, where possibly sub-optimal actions are taken if their EVSI is high enough. However, ARL has some features that distinguish it from standard RL:

1. In multi-armed bandits, the agent should only explore arms believed to be sub-optimal if they will query them.
2. In MDPs, an agent with knowledge of the transition function can avoid querying certain state-actions all together. Some state-actions are *unavoidable* on any policy. For example, the starting board in Chess is visited exactly once per game.
3. The agent can learn about the transition function P without paying any query cost. This distinguishes model-learning from reward-learning. There are situations where the optimal agent learns about P first (before learning anything about R) in order to direct subsequent queries more intelligently.

3.1 Algorithms

We consider model-based algorithms for ARL with episodic, tabular MDPs. Our models are based on *Posterior Sampling for Reinforcement Learning* (Strens, 2000), an algorithm for episodic MDPs that is state-of-the-art in its empirical performance and its known regret bounds (Osband and Van Roy, 2016). As with PSRL, our algorithms assume the agent has a Bayesian probability distribution on the reward function which is updated every episode. We follow PSRL in planning according to a reward function sampled from the posterior distribution (i.e. Thomson-sampling). The posterior distribution will also be used to estimate the value of querying.

An obvious baseline is simply to use PSRL and query each state-action the first N times it is visited. The question is how to choose the hyperparameter N . Our first algorithm provides one approach.

3.1.1 Simulating Query Strategies with Monte Carlo Rollouts

The *Simulated Query Rollout (SQR)* algorithm computes the average performance of a given query strategy (e.g. a value for hyperparameter N) in environments sampled from the agent’s posterior. This involves simulating the agent’s performance across all the episodes for many different environments. This limits its use to choosing between a small set of possible query strategies. In our experiments, we use SQR to tune N , the number of times to query each state-action. We also consider an approximate version of SQR (*ASQR*), which assumes the agent learns about queried rewards directly (rather than having to actually visit states to query them). By running ASQR at the start of every episode (*ASQR in the loop*), the agent picks an N tuned to the remaining time (rather than a fixed N for the whole run).

3.1.2 Estimating the Value of Information

The algorithms above were based on the idea of picking an N and querying each state-action N times. Yet some state-actions are unavoidable (or hard-to-avoid on any policy) and these should never be queried.

Motivated by this observation about unavoidable states, we present algorithms based on approximating the Value of Information of each state-action (s, a) independently. To estimate the value of knowing the reward $R(s, a)$, we compare the performance of ignorant vs. informed agents. The ignorant and informed agents may be, respectively:

1. The current agent, and the current agent with additional knowledge of $R(s, a)$ (“Greedy VOI”).
2. The agent which knows the value of R everywhere except at (s, a) (where it uses the current agent’s beliefs), and the fully omniscient agent (“Omniscient VOI”).

We use the following greedy estimate of the value of information (*VOI*) which does not account for the possibility of further information gathering:

$$VOI = \mathbb{E}_{P(\mathcal{M})}[V_{\mathcal{M}}(\pi_{inf}) - V_{\mathcal{M}}(\pi_{ign})] \quad (2)$$

where $P(\mathcal{M})$ is the agent’s prior over environments, $V_{\mathcal{M}}(\pi)$ is the expected returns of policy π in environment \mathcal{M} , and π_{inf}, π_{ign} are the informed and ignorant agents’ policies, respectively (in our case, the optimal policies in the informed / ignorant agents’ expected environments).

Our algorithm computes $N(s, a)$ as a function of the *VOI* of $R(s, a)$, estimated using Greedy VOI or Omniscient VOI, as well as the number of episodes remaining, E , the query cost, c , and a hyper-parameter, k , which controls the agent’s eagerness to query:

$$N(s, a) = \frac{k \cdot E \cdot VOI[R(s, a)]}{c} \quad (3)$$

Note that we use the value of *perfect* information (EVPI), i.e. knowing $R(s, a)$, not of *sample* information (EVSI), i.e. observing some finite number of rewards sampled from $R(s, a)$, which we plan to explore next. Besides using EVSI, these algorithms should also be extended to account for the opportunity cost of seeking to query a given state, which may be high if that state is hard to access, e.g. due to stochasticity in the environment.

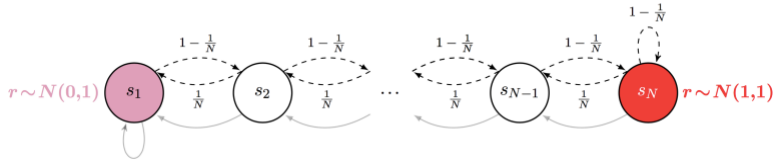


Figure 3: The chain environment used in Osband and Van Roy (2016); our version has deterministic transitions.

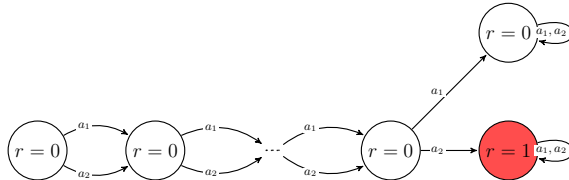


Figure 4: The long-Y environment. Ideally, the agent should only query the two rightmost states, since the other states are unavoidable.

3.2 Experiments

We benchmark our proposed algorithms in the following 3 environments:

1. A chain of length 10, with reward only at the end (see Figure 3).
2. A novel "long-Y" environment, also of length 10 (see Figure 4). This environment has many unavoidable states not worth querying.
3. A 4x4 gridworld with rewards of 0, 1/3, 2/3, 1 along the diagonal and 0 everywhere else.

We compare for small ($c = 1$) and large ($c = 10$) query costs, and run for 4096 episodes⁴. The agent has an independent standard normal prior for the reward of each state (or state-action, in the case of the chain) and learns only the mean (the variance is fixed to 1). This prior is somewhat optimistic given the true reward structure.

We resample environments each episode for PSRL and to update our query strategy. We use a single sampled environment for the Monte Carlo estimate of performance, although using more samples can significantly increase performance. For VOI algorithms, we set $k = 1$. For the baseline algorithms, we either query each state-action up to 25 times (baseline 2), or query up to a total of $25|\mathcal{S}||\mathcal{A}|$ times (baseline 1); these numbers were tuned by hand based on previous experiments, which would not generally be possible in real applications.

Our results are presented in Figure 5.

As expected, the VOI algorithms perform better in the long-Y environment, as a result of not querying unavoidable states. In the chain environment, however, the ASQR-based algorithms perform better; results in gridworld (not pictured) were similar. Although the VOI algorithms have higher returns, they also query more, and hence perform worse overall; using smaller values of k might improve their performance. The VOI algorithms also query relatively less often on earlier episodes, which is undesirable. This is because even when knowing $R(s, a)$ is valuable, in many sampled environments, it will not be. Sampling more environments, or using an algorithm based on confidence-bounds would encourage earlier querying. Not querying at all performs remarkably well in the chain environment when the query cost is high, likely due to the difficulty of the task and the optimism of the agent's reward prior. Overall, these results demonstrate the value of updating query strategies online, and of differentially querying different state-actions, but should not be taken as a rigorous evaluation of the strength of the proposed algorithms.

⁴ Recall that the agent knows the number of episodes and uses this information to inform its query strategy.

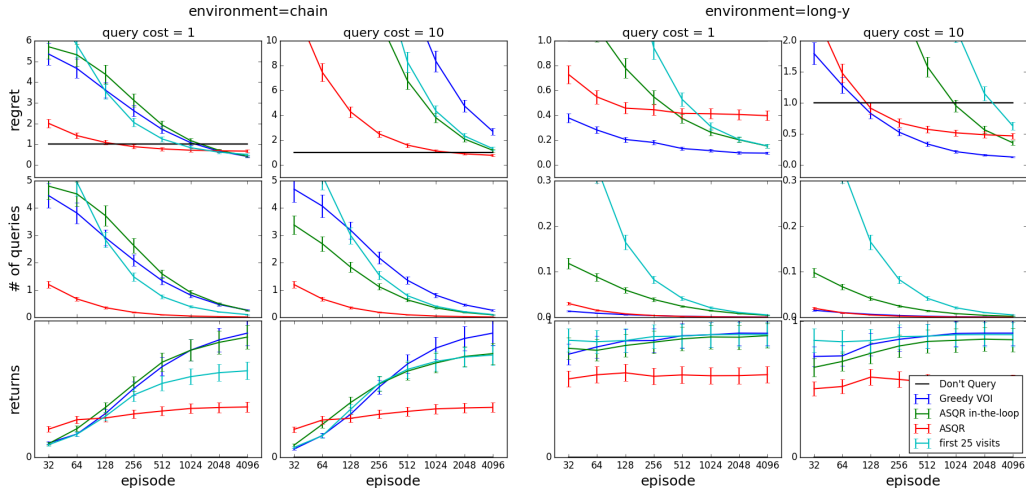


Figure 5: Average cumulative regret per episode (top), average number of queries per episode (middle), and average returns per episode (bottom) for different ARL algorithms in chain (left columns) and long-y (right columns) environments, with standard error bars.

4 Conclusion

We motivated and described Active Reinforcement Learning (ARL), in which an agent must pay to observe its reward signal. We demonstrated important properties that differentiate ARL from standard RL in both the multi-armed bandit and MDP cases. We explored a range of heuristic algorithms for these problems, some of which obtained promising empirical results (although our experiments were limited in scope).

Although we focused on the model-based case, where estimating performance of different query strategies is straightforward, we intend in future work to examine model-free approaches to ARL. Estimating the value of information or learning not to query unavoidable states seems challenging without a model, although visit counts could provide some relevant information. Our current work focuses on multi-armed bandits and tabular MDPs with known dynamics. Future work will investigate MDPs with unknown dynamics and large state spaces (for which function approximate is necessary).

Acknowledgments. This work was supported by Future of Life Institute grant 2015-144846 (JS, DK, OE). We thank Andreas Stuhlmüller, Tor Lattimore, Reimar Leike, Ryan Lowe, Akram Erraqabi, and Jessica Taylor for helpful discussions. The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work.

References

- András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.
- Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.
- C. Daniel, M. Viering, J. Metz, O. Kroemer, and J. Peters. Active reward learning. In *Robotics: Science & Systems*, 2014.
- Owain Evans, Andreas Stuhlmüller, and Noah D Goodman. Learning the preferences of ignorant, inconsistent agents. *arXiv preprint arXiv:1512.05832*, 2015.
- Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. Selecting computations: Theory and applications. In *Uncertainty in Artificial Intelligence*, 2012.

- Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Conference on Learning Theory*, 2010.
- W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The TAMER framework. In *International Conference on Knowledge Capture*. ACM, 2009.
- Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. In *Neural Information Processing Systems*, 2015.
- Manuel Lopes and Luis Montesano. Active learning for autonomous intelligent agents: Exploration, curiosity, and interaction. *arXiv preprint arXiv:1403.1497*, 2014.
- Omid Madani, Daniel J Lizotte, and Russell Greiner. The budgeted multi-armed bandit problem. In *Computational Learning Theory*, 2004.
- Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.
- Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning. *arXiv preprint arXiv:1607.00215*, 2016.
- Billy Okal Philip S. Thomas. A notation for Markov decision processes. *CoRR*, abs/1512.09075, 2015.
- Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Computational Learning Theory*, 2001.
- Warren Powell and Ilya Ryzhov. *Optimal Learning*. John Wiley & Sons, 2012.
- Howard Raiffa and Robert Schlaifer. Applied statistical decision theory, 1961. URL <http://opac.inria.fr/record=b1082847>.
- Kevin Regan and Craig Boutilier. Robust online optimization of reward-uncertain MDPs. In *International Joint Conference on Artificial Intelligence*, 2011.
- Malcolm Strens. A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, 2000.
- Paul Weng, Robert Busa-Fekete, and Eyke Hüllermeier. Interactive Q-learning with ordinal rewards and unreliable tutor. In *The ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*, 2013.
- Christian Wirth and Johannes Fürnkranz. Preference-based reinforcement learning: A preliminary survey. In *The ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*, 2013.