Iterative Interactive Reward Learning

Pallavi Koppol [1] **Henny Admoni** [2] **Reid Simmons** [1 2]

Abstract

Machine learning systems are frequently interacting with people in changeable environments, and would benefit from being able to leverage insights from those people. There are a number of interaction types that these increasingly sophisticated systems can use towards this end. However, people have limitations that affect their teaching and need to be accounted for. One such limitation is a finite working memory capacity. We propose that interaction types that are more informative also result in increased cognitive load. Thus, we present a design for a learning framework that iterates between demonstration, rating, and preference interactions in order to preserve learning performance while minimizing human cognitive effort.

1. Introduction

People interact with disembodied systems such as recommendation engines on a near-daily basis, and embodied systems such as self-driving cars promise to soon become fixtures in daily life. Alongside this omnipresence comes the need for these systems to behave appropriately in an unprecedented number of possible contexts. To support such breadth of behavior, these systems can leverage human insight.

Currently, most people who interact with these systems play no role in shaping their development; thus, a wealth of insights, preferences, and priorities go unheard. For example, should a self-driving car get stuck in a situation where it does not know how to act appropriately (e.g. merging safely, or navigating a crowded parking lot), its passenger can likely identify a desirable course of behavior. However, it is unclear how they could teach this complex system. The difficulty in teaching complex behaviors is consistent throughout the learning process, and without regard for whether the instructor is a layperson or a machine learning expert. The question then becomes: how can we build learning systems that are easier for people to teach?

We concern ourselves with techniques for learning reward functions in order to derive optimal behavioral policies for an intelligent agent, such as a self-driving car. Researchers have investigated a variety of interaction types that can be leveraged towards this end. These include, but are not limited to, learning from human-provided demonstrations, critiques, ratings, corrections, and preferences (Abbeel & Ng, 2004; Cui & Niekum, 2018; Daniel et al., 2015; Bajcsy et al., 2018; Dorsa Sadigh et al., 2017). The diversity in available techniques may stem in part from their complementary natures, and the varied benefits each technique brings to the table. For example, demonstrations are more informative than preferences, but less accurate, and both can be combined to accelerate learning (Palan et al., 2019). Work from (Bullard et al., 2019) has also explored game-theoretic and heuristic based approaches to switching between interaction types as an avenue towards increased learning efficiency; this work also assumed a constrained query budget in order to approximate the limitations of a human instructor.

Taking these types of limitations into consideration is necessary because people are flawed teachers. Unlike oracle agents, they are noisy, they are overly generous in their feedback, and they get fatigued by long streams of questions (Amershi et al., 2014). Several of the shortcomings in people's teaching capabilities may relate to their limited working memory capacity (Miller, 1956). As the cognitive load (i.e. the portion of working memory being utilized) on an individual increases, they grow more easily distracted and tend to have worse task performance (Sweller, 1988). Interestingly, interaction design can be used to modulate cognitive load in human learners (Chandler & Sweller, 1991); that is, the way a task is presented can affect how burdensome it is to complete it.

We propose that *interaction types that are more informative also result in increased cognitive load*. Given this proposal, we can construct a learning framework that dynamically iterates between using demonstration, rating, and preference interactions in order to preserve learning performance while minimizing human cognitive-effort.

The main contributions of this paper are the motivation for

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA [2]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Pallavi Koppol <pkoppol@andrew.cmu.edu>.

and design of such an iterative, interactive reward learning system. In addition, we present preliminary thoughts on evaluations and directions for further exploration.

2. Related Work

2.1. Learning from People

We are particularly interested in how to learn a reward function from a person. Reward functions have been described as the most compact and transferable descriptor of a task (Russell, 1998), and can capture a person's needs, expectations, and preferences. We focus on the literature regarding learning rewards from demonstrations, ratings, and preferences.

Learning from Demonstration: Without any prior information, the space of a user's potential reward functions may be expansive. Demonstrations can be used to narrow down this space. Inverse Reinforcement Learning (IRL) is a technique that is used precisely for recovering a reward function given a series of demonstrations (Ng et al., 2000). Commonly, the reward function can be modeled via a linear combination of feature weights (Abbeel & Ng, 2004); this is an assumption that we make as well. IRL can also be formulated from a Bayesian perspective (BIRL), wherein the demonstrator's actions serve as evidence that is used to update a prior distribution over reward functions (Ramachandran & Amir, 2007). In another Bayesian approach, (Levine et al., 2011) explored how Gaussian Process (GP) can be used for inverse reinforcement learning. We similarly use a GP to explicitly maintain a distribution over feature weights that other interaction types can leverage as well.

While demonstrations can be highly informative, they are also difficult to provide. Due to time and resource constraints, people are limited in the number of examples they can provide, particularly as the task grows more complex. Furthermore, IRL is known to suffer from a degeneracy problem: multiple rewards can explain a single behavior (Ng et al., 2000). Finally, people may not be able to demonstrate what they would like but rather what they are capable of (Basu et al., 2017).

Active Reward Learning with Labels: Active learners strive to be data efficient by finding the most informative query at each iteration. A common approach is to utilize active labeling, wherein a query point is selected and a user must ascribe a label to it. GPs are particularly well suited to this, and there are a host of well-developed Bayesian Optimization techniques and acquisition functions that researchers have explored in the context of reward learning (Daniel et al., 2015). In our work, we utilize an *active ratings* technique, which is a subset of active labeling where users are given a discrete five-point scale along which to rate queries.

While it has been found that label queries are easier to answer than demonstration queries (Cakmak & Thomaz, 2012), a discrete scale restricts the informativeness of any answer. Furthermore, people are known to be overly generous in their feedback, which might skew results (Amershi et al., 2014). Additionally, people's ratings are not static: they tend to increase, meaning that a query rated low earlier on in the training process might be given a higher score later as the user's expectations adjust (O'Connor & Cheema, 2018). Therefore, they are well-suited to be used in conjunction with other interaction types.

Active Preference-based Reward Learning: There is a growing body of work on comparison-based approaches to learning reward functions. The goal of the Active Preference-based Reward Learning (APbRL) technique presented in (Dorsa Sadigh et al., 2017; Erdem et al., 2020) is to recover a user's underlying reward function, given as a linear combination of weights, for a trajectory planning task by presenting users with a series of comparison queries. In (Bıyık et al., 2020), the reward function is modeled using a GP. We use a similar GP formulation in our work.

Preference queries are precise, require little effort from users, and consist of choosing between two (or more) potential options. As a result, the technique is good at fine-tuning a coarse understanding of a user's reward function, but the amount of information that can be gained from each query is limited. Furthermore, this technique can be inefficient due to the computational overhead of repeatedly optimizing candidate preference queries online.

2.2. Mixing Interaction Types

It is known that people utilize multiple interaction types when learning a novel task. (Cakmak & Thomaz, 2012) categorized people's questions into three primary types: labels, demonstrations, and feature queries. Researchers have explored how multiple interaction types can be leveraged by algorithmic learners as well. (Palan et al., 2019) recognized that demonstrations and preferences traded off informativeness with accuracy, and that combining them reduced convergence time: demonstrations could be used to narrow down a user's potential reward space, and preferences to hone in on the true reward function. Further work has investigated how to toggle between interaction types when the agent has a limited number of queries it can make: game-theoretic, rule-based, and learned questioning strategies were explored in (Bullard et al., 2018; 2019). Our work also explores how multiple interaction types can be used in conjunction, though we use the cognitive load induced by each one as a key parameter in selecting which to use.

3. Problem Definition

Goal: Our system will utilize various types of user feedback in order to recover a reward function that leads to desirable agent behavior.

Model: We consider a sequential decision making problem with states $s \in S$ and actions $a \in A$, that is both deterministic and fully observable.

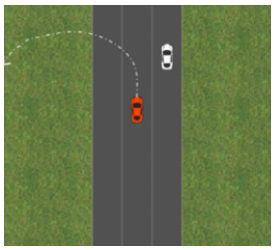

Figure 1. The dashed line in this figure represents a trajectory ξ that might be displayed to a user, who would then have the option of returning a rating from 1 (worst) to 5 (best).

As in (Erdem et al., 2020), the pair (s_t, a_t) denotes the state of the environment at time t and the corresponding action taken. A trajectory represents a series of these pairs, and can be defined as $\xi = ((s_t, a_t))_{t=0}^{T}$, where T is a finite time horizon. Figure 1 shows an example trajectory in our self-driving car domain.

We make the standard assumption that the reward is given by a linear combination of feature weights (Abbeel & Ng, 2004), which reduces the reward-learning problem to learning feature weights. That is, $R(\xi) = w^T \phi(\xi)$ where $w \in \mathbb{R}^d$ represents the weights we need to learn, and $\phi(\xi) \in \mathbb{R}^d$ is a feature function that evaluates feature values over ξ. We constrain $||w||_2 = 1$ without loss of generality as in (Brown & Niekum, 2018; Palan et al., 2019). Initially, we assume a uniform prior over w, which is updated based on user feedback.

Interaction Types: We leverage three interaction types in service of our goal.

Demonstration Queries ($Q_D = s_0$), prompt the user with an initial state and request the provision of a desirable trajectory in the provided environment.

Rating Queries ($Q_R = \xi$) prompt the user with a single trajectory and request a rating on a discrete five-point scale.

Preference Queries ($Q_P = (\xi_A, \xi_B)$), prompt the user with two trajectories, and request the selection of the superior trajectory.

Having a unified framework will allow us to toggle between different interaction types at each querying iteration. We leverage a standard multi-variate GP with a radial basis function kernel in order to maintain a distribution over the learned reward function and unify the aforementioned interaction types. Utilizing a GP framework allows us to flexibly make use of Bayesian Optimization techniques, and explicitly reason about uncertainty over the distribution. We build on pre-existing approaches towards GP-based learning from demonstration and preferences (Levine et al., 2011; Bıyık et al., 2020), and implement our own ratings technique using a discrete five-point scale and a traditional Upper Confidence Bound acquisition function (Srinivas et al., 2009).

Informativeness and Cognitive Load: The previously defined interaction types have a trade-off between informativeness and cognitive load on the user. Our approach seeks to utilize the most informative and effortless interaction type at each querying step. Thus, we need to formalize definitions for informativeness and cognitive load that can be used in our optimization procedure.

Let $\rho_x(w)$ denote the informativeness of interaction type $x \in \{d, r, p\}$, where w is the current distribution over the reward space. Similarly, let $c_x(i)$ be the cognitive load of interaction type x at iteration i of the learning system. Note that neither $\rho_x(w)$ nor $c_x(i)$ are static variables: the informativeness of a query is affected by the shape of the distribution that query is drawn from, and cognitive load is likely to increase with task length.

Our goal at each iteration i is to solve the following optimization for the ideal interaction type x^*:

$$x^* = \arg\max_{x \in \{d,r,p\}} \alpha \rho_x(w) - (1-\alpha) c_x(i) \tag{1}$$

Here, $\alpha \in [0, 1]$ is a tuning parameter. Initially, let $\alpha = 0.5$.

Each time we query a user, we want to use the interaction type that maximizes the amount of information gained while minimizing induced cognitive load.

To do this, we need to prescribe real-values to the informativeness and associated cognitive load of each interaction type. We propose using the expected information-gain of a particular interaction type as an initial definition of informativeness, as in (Jeon et al., 2020; Erdem et al., 2020):

$$\rho_x(w) = I(q_x; w | Q_x) \tag{2}$$

$$= \mathbb{E}_{w, q_x | Q_x} \left[\log \left(\frac{P(q_x | w, Q_x)}{\int P(q_x | w', Q_x) P(Q_x | w') dw'} \right) \right] \tag{3}$$

In Eq.(2), q_x is the response from the user (e.g. their demonstration, rating, or preference selection) to the query Q_x, and the integral must be taken over the entire space of w. Precisely computing $P(q_x | w, Q_x)$ for $x \in \{d, r, p\}$ is out-

side the scope of this paper, and will be addressed in future work.

Furthermore, in order to determine the cognitive load of each interaction type, we need to run a user study and collect data. We would need to run a study wherein participants use each interaction type sequentially (e.g. a section of demonstration queries, followed by ratings queries, followed by preference queries). We could then collect subjective measures of cognitive load, such as the widely used NASA-TLX questionnaire, which allows us to compute a task load score between 0 and 100 (Hart & Staveland, 1988). Simultaneously, we could use this data to confirm our informativeness metrics.

Let $s_x, x \in \{d, r, p\}$ denote the task load score obtained through the study. Then, let $c_x(i) = k(i) \cdot s_x$, where $k(i)$ is the factor by which cognitive load increase with each query iteration. We leave finding the optimal fatigue formula $k(i)$ for future work; as a first pass, we can let $k(i) = 1$. With this, we now have enough information to solve Eq. (1) and determine which interaction type to use.

4. Proposed Evaluation

Our learning framework will be evaluated along the two axes of learning performance and user-friendliness. We seek to answer questions including (1) *How well, and how quickly, can our system recover a ground truth reward function?* (2) *How satisfied are users with the final performance of the system?* (3) *Would a user utilize this system to teach an agent again?* and (4) *What is the overall cognitive effort induced by this system?* While (1) can be answered via the use of a simulated oracle user, we need to run a user study in order to answer the remainder. Ultimately, we want to understand the answers to these questions in relation to other state-of-the art techniques.

These questions can be answered using both objective and subjective task performance measures. Objective measures might include the time taken for a user to respond to a particular interaction type, or their ability to perform a secondary task alongside responding to queries. Subjective measures include responses to survey questions regarding the user's experience with the system or their desire to use such a system again, as well as measures of cognitive load such as the NASA-TLX questionnaire.

There are a number of baselines against which we can compare our method in order to collect the aforementioned metrics. For example, we can compare our system against demonstration-only, ratings-only, and preference-only active learning systems in order to understand the efficacy of multiple interaction types on learning performance and user-friendliness. In order to understand the efficacy of iterative querying, we can compare this system to a unified approach with a fixed ordering of interaction types. Ultimately, we will run user studies with baselines and metrics that comprehensively evaluate the usability of our proposed learning framework.

5. Conclusion

In this paper, we proposed a design for an iterative, interactive reward learning system. We designed this system to intelligently toggle between interaction types in order to minimize the cognitive load on a user without reduction in learning performance. We outlined potential user studies to further elucidate this trade-off between informativeness and cognitive load, and evaluate our system's approach. This work is but one component of a larger effort to actively accommodate human guidance in machine learning.

Though this project is a work in progress, we anticipate that there will be several promising directions for future exploration. For example, human teachers typically look for and respond to signs of growth and learning from their students, and respond positively to transparency in the machine learning process (Amershi et al., 2014). With this in mind, we might investigate how a unified GP framework can enable more explicit two-way communication between teacher and learner. Alternatively, we can study how user expertise affects both teaching and learning performance.

6. Acknowledgements

This work is supported in part by the Office of Naval Research (N00014-18-1-2503).

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1. ACM, 2004.

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

Bajcsy, A., Losey, D. P., O'Malley, M. K., and Dragan, A. D. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 141–149. ACM, 2018.

Basu, C., Yang, Q., Hungerman, D., Sinahal, M., and Draqan, A. D. Do you want your autonomous car to drive like you? In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 417–425. IEEE, 2017.

Bıyık, E., Huynh, N., Kochenderfer, M. J., and Sadigh, D. Active preference-based gaussian process regression for reward learning. 2020.

Brown, D. S. and Niekum, S. Efficient probabilistic performance bounds for inverse reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Bullard, K., Thomaz, A. L., and Chernova, S. Towards intelligent arbitration of diverse active learning queries. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6049–6056. IEEE, 2018.

Bullard, K., Schroecker, Y., and Chernova, S. Active learning within constrained environments through imitation of an expert questioner. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Cakmak, M. and Thomaz, A. L. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 17–24. ACM, 2012.

Chandler, P. and Sweller, J. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4): 293–332, 1991.

Cui, Y. and Niekum, S. Active reward learning from critiques. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6907–6914. IEEE, 2018.

Daniel, C., Kroemer, O., Viering, M., Metz, J., and Peters, J. Active reward learning with a novel acquisition function. *Autonomous Robots*, 39(3):389–405, 2015.

Dorsa Sadigh, A. D. D., Sastry, S., and Seshia, S. A. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

Erdem, B., Palan, M., Landolfi, N. C., Losey, D. P., Sadigh, D., et al. Asking easy questions: A user-friendly approach to active reward learning. In *Conference on Robot Learning*, pp. 1177–1190, 2020.

Hart, S. G. and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pp. 139–183. Elsevier, 1988.

Jeon, H. J., Milli, S., and Dragan, A. D. Reward-rational (implicit) choice: A unifying formalism for reward learning. *arXiv preprint arXiv:2002.04833*, 2020.

Levine, S., Popovic, Z., and Koltun, V. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 19–27, 2011.

Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Ng, A. Y., Russell, S. J., et al. Algorithms for inverse reinforcement learning. In *ICML*, volume 1, pp. 2, 2000.

O'Connor, K. and Cheema, A. Do evaluations rise with experience? *Psychological Science*, 29(5):779–790, 2018.

Palan, M., Landolfi, N. C., Shevchuk, G., and Sadigh, D. Learning reward functions by integrating human demonstrations and preferences. In *Robotics: Science and Systems (RSS)*, 2019.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

Russell, S. J. Learning agents for uncertain environments. In *COLT*, volume 98, pp. 101–103, 1998.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Sweller, J. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.