Real-Time Interactive Reinforcement Learning for Robots

Andrea Lockerd Thomaz, Guy Hoffman, and Cynthia Breazeal
Robotic Life Group
MIT Media Lab
20 Ames Street
Cambridge, MA 02139
{alockerd,guy,cynthiab}@media.mit.edu

Abstract

It is our goal to understand the role real-time human interaction can play in machine learning algorithms for robots. In this paper we present *Interactive Reinforcement Learning* (IRL) as a plausible approach for training human-centric assistive robots by natural interaction. We describe an experimental platform to study IRL, pose questions arising from IRL, and discuss initial observations obtained during the development of our system.

Introduction

Machine learning is bound to play a significant role in the development of robotic assistants for human environments such as homes, schools, hospitals, nursing homes, and offices. Taking into account the complexity and unpredictability of both the surroundings and the tasks that such robots will have to accommodate for, we expect that a significant amount of training will be carried out by (possibly unskilled) humans on-location. In order for assistive robots to play a long-term role in people's lives, we should additionally expect the need for the robot to periodically learn new tasks and task refinements.

While various works have acknowledged and addressed some of the hard problems that robots face when learning in the real-world (Mataric 1997; Thrun & Mitchell 1993; Thrun 2002), social learning in a human environment, as described above, poses additional challenges for machine learning systems. Autonomous robots will need to learn from *natural* interactions with untrained humans. Given the limits of human attention and patience, they will need to learn in real-time from relatively few examples. Since every action utilizes precious time and resources (battery, motor life, etc.) each trial run must be thoroughly exploited. The tolerance for trials ending in failure is bounded at best (and often nonexistent) and realistically is far below the vast number usually relied upon for learning via exploration. In summary, human-centric robots will have to learn efficiently, safely, and from few examples.

Interactive Reinforcement Learning

To this end we are developing the idea of *Interactive Reinforcement Learning* (IRL) for autonomous robotic agents. We use the term "IRL" as a generic name for a variation on traditional Reinforcement Learning (RL) algorithms, in which the reward signal is determined not only based on the world state and agent action, but additionally by a real-time interaction with a human teacher or coach. In an IRL session, the human can choose to change the reward signal not only at certain goal states, but continuously throughout the interaction. However, IRL also allows the human to remain passive when appropriate, usually deferring to an independent environmental based reward signal.

Understanding the impact that human interaction has on a standard machine learning process is the overarching goal of this research. Our work on this problem domain is in its preliminary stages. In this paper we present an experimental platform to study IRL, pose some questions we are working to resolve, discuss initial observations obtained during the development of our system, and project possible ways to address the new challenges presented by IRL for human-centric robots.

Research Questions

We want to study how human interaction can and should change the machine learning process. How does natural human feedback work with standard reinforcement learning algorithms? What are the new constraints? What behavioral feedback does the human teacher need in order to understand the process? Motivated by these questions and others, we are developing a simulation environment that can serve as a tool for experimenting with various algorithms and enhancements. The ultimate goal of this research is to use these resulting learning mechanisms on a humanoid robot.

IRL emphasizes the *interactive* elements in teaching. There are inherently two sides to an interaction, in our case the teacher and the learner. Our approach aims to enhance standard machine learning algorithms from both perspectives of this interaction.

How Do Humans Teach?

Significant strides have been made in the development of machine learning algorithms in structured environments

(static and dynamic). However, in order to adapt these machine learning methods for robots that are to learn from real-time natural human instruction, we must take into account factors of human interaction. An important question for IRL is "how do humans want to teach?". Our approach begins with a few hypotheses about this question. We expect to learn about these issues through experiments with real people.

One aspect of human teaching is the *timing* of feedback given to the learning agent. While timing of reward signals, and particularly dealing with delayed rewards have been discussed in RL literature (Kaelbling, Littman, & Moore 1996), the timing of an untrained human's feedback in an IRL situation has received little attention, yielding numerous open questions:

- Is there a lower bound on time between rewards?
- Does timing change at different stages of the interaction?
- Which of the agent's "behavioral" elements influence the timing of the feedback?
- How does habituation influence reward-giving?

A second human factor is the *level* of feedback that a human teacher would want to give an interactive machine learning agent.

- When does feedback pertain to the whole task versus a specific action in the task?
- How often is feedback associated with a state as a whole, as opposed to a certain aspect of a state?
- What agent behavior can elicit high-level steering as opposed to micro-management of the agent's actions?

A meta-issue involved in social learning with a human teacher is that interactions with different teachers could be quite different.

- Are there different teaching styles, or coach types?
- Can these be automatically detected?
- Does it make sense to adapt the agent's behavior to a specific teaching pattern?

We believe that all of the above will prove crucial for designing machine learning algorithms for robots that are to learn in a natural human environment.

Transparency - Revealing the Internal State

The above questions pertain to factors in the human teacher's behavior. Another important question is how the learning agent's behavior can affect the learning process. A social learning interaction is inherently a partnership, where both learner and teacher influence the performance of the tutorial dyad. While this observation seems straightforward, little attention has been paid to the communication between human teacher and artificial agent in the traditional machine learning literature.

Particularly, we believe that the *transparency* of the learner's internal process is paramount to the success of the tutorial dialog. However, a fine balance must be struck between engulfing the human teacher with all pertinent information, and leaving them in the dark.

A central goal of the presented architecture is the investigation of transparency mechanisms that are intuitive for untrained human coaches of machine learning robotic agents. We look towards early child development literature to inform our choice of agent-to-teacher signals, and therefore stress the use of *gesture*, *gaze*, and *hesitation* as intuitively readable signals for an IRL agent (Kaye 1977; Krauss, Chen, & Chawla 1996; Argyle, Ingham, & McCallin 1973)

Experimental Platform: Sophie's World

To investigate the above questions regarding IRL, we have implemented a Java-based simulation platform, which we dub *"Sophie's World"* (SW). SW is a generic object-based State-Action MDP space for a single agent, Sophie, using a fixed set of actions on a fixed set of objects. The semantics in Sophie's World are generic with respect to task context or learning algorithm, enabling it to be used with a variety of specific tasks and learning modules.

The design of SW was done with modularity and web-deployment in mind. We plan to substitute various learning algorithms and utilize different task scenarios, while keeping the human interaction elements unchanged. We also plan to deploy SW as a web-based application, enabling the collection of a large amount of naïve-user data towards our goal of understanding how human teachers steer a learning agent.

Sophie's MDP

In our system, a World $W = \langle L, O, \Sigma, T \rangle$ is defined by a finite set of k locations $L = \{l_1, \ldots, l_k\}$ and n objects $O = \{o_1, \ldots, o_n\}$. Each object can be in one of an object-specific number of mutually exclusive object states. If we denote the set of states object o_i can be in as Ω_i, then the complete object configuration space can be described as $O^* = \langle \Omega_1 \times \ldots \times \Omega_n \rangle$.

W is also defined by a set of legal states $\Sigma \subset \langle L \times L^O \times O^* \rangle$. Thus a world state $s(l_a, l_{o_1} \ldots l_{o_n}, \omega)$ consists of the location of the agent, the location of each of the objects, and the object configuration in $\omega \in O^*$.

Finally, W has a transition function $T : \Sigma \times A \mapsto \Sigma$, where A is fixed (see below).

The action space A is defined by four predefined atomic actions with arguments as follows: Assuming the locations L are arranged in a ring, at any time step, the agent can `GO left` or `right`, moving to a different location in the ring; she can `PICK-UP` any object that is in her current location; she can `PUT-DOWN` any object currently in her possession; and she can `USE` any object in her possession on any object in her current location. Each action implements a transition function in T that advances the world state.

On top of the described generic architecture we can then build task-specific implementations, which determine the remaining aspects of the world: the the spatial relationship between the locations, the limit on the number of objects in the agent's possession, and the sub-transition $T_U \subset T$ that is accomplished by the `USE` action.

Interactive Rewards Interface

A central feature of SW is the interactive reward interface. Using the mouse, a human trainer can—at any point in the operation of the agent—award a scalar reward signal $r = [-1, 1]$. The user receives visual feedback enabling them to tune the reward signal to a certain value before sending it to the agent. Choosing and sending the reward value does not halt the progress of the agent, which runs asynchronously to the interactive human reward.

Learning Algorithms

We believe that the reinforcement learning paradigm lends itself naturally to a human interaction component, even though there are few examples of interactive RL. Initially the algorithms we are implementing include a standard Q-Learning algorithm (Russell & Norvig 2002), a hierarchical SMDP approach (Sutton, Precup, & Singh 1998), and we are also looking at a recent biologically inspired form of hierarchical RL (Singh, Barto, & Chentanez 2005).

We view this as a platform to explore and compare multiple learning algorithms to understand how they can (or cannot) be used in an interactive setting with untrained users. The goal is to understand, under different approaches, how human interaction and transparent expression of internal state can be incorporated into the learning process.

Transparency behaviors

Adding elements of internal state transparency is expected to significantly alter the learning process. The first element of transparency we are exploring is gaze. The use of gaze requires that the learning agent have a physical embodiment (either real or virtual) that can be viewed and understood by the human as having a forward heading or in some way orienting prior to action. Thus, gaze precedes each action and communicates something about the action that is going to follow. More abstract functions of gaze that we also hope to explore include gaze as an expression of awareness (looking at something establishes the mutual knowledge that you know about it) and gaze as an expression of uncertainty (looking between two potential action targets can communicate low confidence).

The Kitchen World

To exemplify a usage of SW, our first implemented task scenario is a Kitchen world (see Fig 1). In it the agent is to learn to prepare batter for a cake and put the batter in the oven.

The object space contains five objects: `Flour`, `Eggs`, a `Spoon` (each with a single object state), a `Bowl` (with five object states: `empty`, `flour`, `eggs`, `unstirred`, and `stirred`), and a `Tray` (with three object states: `empty`, `batter`, and `baked`). The world has four locations: `Shelf`, `Table`, `Oven`, and `Agent` (i.e., the agent is in the center with a shelf, table and oven surrounding her).

In this kitchen implementation, the agent can hold at most one thing at a time. `USE`ing an ingredient on the `Bowl` puts that ingredient in it; using the `Spoon` on the `unstirred Bowl` transitions its internal state to `stirred`; Using the `stirred Bowl` on an `empty Tray` fills it, and so on.

Figure 1: Sophie's Kitchen. The display at the top left indicates the current state-based reward. The vertical bar is the interactive reward interface and is controlled by the human.

In the initial state, S_0, all objects are on the `Shelf`, and the agent faces the `Shelf`.

A successful completion of the task will include putting flour and eggs in the bowl, stirring the ingredients using the spoon, then transferring the batter into the tray, and finally putting the tray in the oven.

Due to the large number of same-state transitions and the flexibility of the state space, there are many action sequences that can lead to the desired goal. Here is one such sequence: `PICK-UP Bowl`; `GO right`; `PUT-DOWN Bowl`; `GO left`; `PICK-UP Flour`; `GO right`; `USE Flour`; `PUT-DOWN Flour`; `GO left`; `PICK-UP Eggs`; `GO right`; `USE Eggs`; `PUT-DOWN Eggs`; `GO left`; `PICK-UP Spoon`; `GO right`; `USE Spoon`; `PUT-DOWN Spoon`; `GO left`; `PICK-UP Tray`; `GO right`; `PUT-DOWN Tray`; `PICK-UP Bowl`; `USE Bowl`; `PUT-DOWN Bowl`; `PICK-UP Tray`; `GO right`; `PUT-DOWN Tray`.

In order to encourage short sequences, an inherent negative reward signal of $\rho = -.04$ is placed in any state but the goal state. Also, some end states are so-called *disaster* states (for example—putting the eggs in the oven), which result in a significant negative reward, the termination of the current trial episode, and a transition to state S_0.

Initial Observations

The following section includes some of our initial observations concerning Interactive Reinforcement Learning gleamed from our own interactions with the system in its current development phase.

It is fairly obvious that a standard Q-Learning agent, for example, is not suitable for learning an assistive task using interactive reward training as described above—if only due to the vast number of trials necessary to form a reasonable policy. However, the details of what exactly needs adjustment, and what human factors are dominant in such an interaction, are largely unexplored. It is these key components

that we wish to uncover and enumerate.

What does the Interactive Reward Pertain To?

The main difference between a standard RL system and the proposed IRL framework is the real-time interactive reward signal that is administered by the human coach. In contrast to a traditional MDP reward signal, where a reward is closely associated with a complete state or a state-action pair, a human-derived reward bears a much more vague relation to the most recent state and action of the agent.

A human coach is expected to naturally communicate in goal-oriented and intentional ways, and it will be up to the agent to determine what the human's communication is *about*. If in most MDP scenarios a reward pertains to a complete state, in an IRL reward there is usually something in particular about the state that is being rewarded (i.e., a particular subset of features that has some meaning or purpose to the task at hand). For example, the human will expect to be able to teach the kitchen agent that it is bad to put the spoon in the fire, no matter what is currently on the table, and they will expect the agent not to continually put the spoon in the fire given small perturbations in the feature space of the state.

Other elements of interactive reward have been noted by (Isbell *et al.* 2001). Specifically their experiments with a reinforcement learning agent interacting with human trainers showed that there is significant variability in the timing of rewards, creating a credit assignment problem, usually associated with delayed reward.

In addition to delayed rewards, we expect that with the addition of transparent feedback (especially revealing preaction gestures like gaze) the human teacher will naturally try to predict the agents intentions, and will administer anticipatory feedback as well. While delayed rewards have been discussed in the Reinforcement Learning community, the concept of rewarding the action you are about to do is novel and will require new tools and attention.

Habituation in Teaching

(Isbell *et al.* 2001) also observed habituation in an interactive teaching task. In the beginning of training users tended to give a great deal of feedback, but after the algorithm becomes sufficiently good the feedback was shown to drop off. The human often overlooks the occasional mistake, and will not reward good behavior. Thus, we might conclude a need for a "no feedback is good feedback" policy, whereby the agent can assume that, in some cases, if the human partner says nothing the agent is doing something right.

There may be ways to battle the teaching habituation through social interaction. If the agent has not received feedback for some time it could simply ask the human partner for encouragement.

Regardless, we must take into account the fact that the human partner has a limited attention span for repetitive teaching. Thus, it will be important for the algorithm to be as efficient as possible. For instance, once a certain part of the state-action space is well known, the agent should probably rush through it to get to the unknown and interesting elements at which it may slow down its action to allow for more feedback from the human instructor. The human partner will likely need to see the progress from their teaching quickly in order to remain motivated to teach.

Beyond Scalar Rewards

As stated, a reward signal is typically stationary and is some function of the environment. It is usually a scalar value indicating positive or negative feedback for being in the current state or for a particular state-action pair. Introducing human-derived real-time reward prompts us to reconsider these assumptions.

For instance, an important characteristic of human social learning interactions is "just-in-time" error correction. The teacher can stop and correct the learner's behavior mid-task. Moreover, when told something is wrong, the obvious next step is to reverse or "undo" the previous action. Standard reinforcement learning algorithms are not designed to change on such a small time scale and symmetry is not typically part of action representation. However, when a human partner is the teacher, the algorithm should act more drastically towards negative feedback, and should try to take actions to revert to the prior state. This leads to broader questions about what "undo" means for various actions. It requires some knowledge about equal and opposite action (e.g. pick-up and put-down), and clearly it is sometimes not possible to undo an action (e.g. "mix the eggs and flour"). This meta-knowledge of one's action space may need to be learned in addition to learning specific tasks.

Finally, we found that a scalar positive/negative feedback signal is a fairly impoverished communication channel for a human-centric tutorial interaction. Having other signals than the scalar reward will likely be helpful to both the human teacher and the machine learning algorithm. For instance, we expect that letting the teacher say "keep going", "that's enough", "try again", etc. should help the algorithm control exploration in ways that will be more readable and reasonable to the human partner.

It might also be worth considering qualitatively different levels of reward. Some things are bad for the task because they are useless, others are bad because they cause irreversible damage to the goal, yet others might not be related to the task, but are inherently bad for the robot's safety.

Related Works

Active learning or learning with queries (Cohn, Ghahramani, & Jordan. 1995) is an approach to machine learning that explicitly acknowledges a human in the loop. In these approaches the human's roles is generally quite constrained, and there is not usually much attention paid to making the workings of the learning process transparent to the human.

Cobot is a software agent that learns how to behave in an online community via Reinforcement Learning, receiving rewards from the members of that community (Isbell *et al.* 2001). While the environment and interaction is quite different than our task goals, the application of a standard reinforcement learning algorithm to an interactive situation with untrained human teachers makes their findings interesting and relevant for this work. Their findings suggest that

the average user did not have behavior that was easy to learn from, reward timing is not regular, and reward values are not always consistent over long periods of time. We hope to follow-up and extend this work with new data from our IRL sessions.

Aspects of social learning are starting to be explored with robots. Natural language has been utilized to frame learning episodes in a classification task (Steels & Kaplan 2001), and for instructing a robot in a navigation task (Lauria *et al.* 2002). In Nicolescu and Matarić (2003), a robot learns a sequential task by following a human demonstration (e.g., learning a path through an environment of objects). Short verbal commands point out information and frame the demonstration; by saying 'bad' the human instructor identifies which part of the robot's task model is incorrect, and the subsequent demonstration is used to correct the problem. In other work we have demonstrated aspects of collaboration and social learning on a humanoid robot, using social cues and gestures to achieve transparency and guide instruction (Lockerd & Breazeal 2004; Breazeal *et al.* 2004).

Conclusions

Even from limited exploration of our experimental platform, we were able to identify a host of new considerations for Reinforcement Learning that stem directly from the introduction of a human-based real-time reward signal. We hope to solidify our observations as the system is further developed and deployed to a larger number of untrained human subjects.

We also believe that our conclusions can extend beyond RL robotic agents to other kinds of feedback-based learning systems. Notions of *about*ness of a reward signal, delayed and anticipatory rewards, and even a generalization of gaze as any forward-looking action can be expected to play a role in any feedback-based machine learning algorithm with a human in the loop.

References

Argyle, M.; Ingham, R.; and McCallin, M. 1973. The different functions of gaze. *Semiotica* 19–32.

Breazeal, C.; Brooks, A.; Gray, J.; Hoffman, G.; Lieberman, J.; Lee, H.; Lockerd, A.; and Mulanda, D. 2004. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics* 1(2).

Cohn, D.; Ghahramani, Z.; and Jordan., M. 1995. Active learning with statistical models. In Tesauro, G.; Touretzky, D.; and Alspector, J., eds., *Advances in Neural Information Processing*, volume 7. Morgan Kaufmann.

Isbell, C.; Shelton, C.; Kearns, M.; Singh, S.; and Stone, P. 2001. Cobot: A social reinforcement learning agent. *5th Intern. Conf. on Autonomous Agents*.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. P. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* 4:237–285.

Kaye, K. 1977. Infants effects upon their mothers teaching strategies. In Glidewell, J., ed., *The Social Context of Learning and Development*. New York: Gardner Press.

Krauss, R. M.; Chen, Y.; and Chawla, P. 1996. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In Zanna, M., ed., *Advances in experimental social psychology*. Tampa: Academic Press. 389–450.

Lauria, S.; Bugmann, G.; Kyriacou, T.; and Klein, E. 2002. Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38(3-4):171–181.

Lockerd, A., and Breazeal, C. 2004. Tutelage and socially guided robot learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Mataric, M. 1997. Reinforcement learning in the multi-robot domain. *Autonomous Robots* 4(1):73–83.

Nicolescu, M. N., and Matarić, M. J. 2003. Natural methods for robot task learning: Instructive demonstrations, generalization and practice. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems*.

Russell, S., and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach (2nd edition)*. Prentice Hall.

Singh, S.; Barto, A. G.; and Chentanez, N. 2005. Intrinsically motivated reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems 17 (NIPS)*.

Steels, L., and Kaplan, F. 2001. Aibo's first words: The social learning of language and meaning. *Evolution of Communication* 4(1):3–32.

Sutton, R.; Precup, D.; and Singh, S. 1998. Between mdps and semi-mdps: Learning, planning and representing knowledge at multiple temporal scales. *Journal of Artificial Intelligence Research* 1:139.

Thrun, S. B., and Mitchell, T. M. 1993. Lifelong robot learning. Technical Report IAI-TR-93-7.

Thrun, S. 2002. Robotics. In Russell, S., and Norvig, P., eds., *Artificial Intelligence: A Modern Approach (2nd edition)*. Prentice Hall.