ROBOT TEACHING FOR HUMAN INVERSE REINFORCEMENT LEARNING

Michael S. Lee
Carnegie Mellon University
ml5@andrew.cmu.edu

Henny Admoni
Carnegie Mellon University
hadmoni@andrew.cmu.edu

Reid Simmons
Carnegie Mellon University
rsimmons@andrew.cmu.edu

ABSTRACT

As robots continue to acquire useful skills, their ability to teach their individual expertise will provide humans the two-fold benefit of 1) learning from robots and 2) co-existing or collaborating fluently with them. For example, robot tutors could help teach handwriting to individual students and delivery robots could convey their navigation preferences to better coordinate with nearby human workers. As humans naturally communicate their behaviors through selective demonstrations, and infer others' through reasoning that resembles inverse reinforcement learning (IRL), we propose a method of teaching that uses demonstrations informative for IRL. But unlike prior work that optimizes solely for IRL, this paper incorporates various human teaching strategies (e.g. scaffolding, simplicity, pattern discovery, testing) to better accommodate human learners. We assess our method with a user study and find favorable results for our use of pattern discovery, simplicity, and testing, but find mixed results for our method of scaffolding that advise clear directions for future work.

1 Introduction

As robots become more capable in tasks once accomplished only by humans, the extent of their influence will depend in part on their ability to teach and convey their skills. From the youngest of us learning to handwrite [1, 2] to experienced practitioners of crafts such as chess, many of us stand to benefit from robots that can effectively teach their mastered skill. Furthermore, our ability to collaborate fluently with robots partly depends on us understanding their behaviors. For example, workers at a construction site could better coordinate with a new delivery robot if the robot could clearly convey its navigation preferences (e.g. when it would choose to go through mud over taking a long detour).

While demonstrations are a natural method of teaching and learning behaviors for humans, its effectiveness still hinges on conveying a set of demonstrations that is informative and comprehensible. As human reasoning over another's decision-making can be partly modeled as IRL (e.g. via Bayesian inference) [3–5], we use inverse reinforcement learning (IRL) to quantify the informativeness of demonstrations. But unlike prior work that optimizes solely for IRL [6], this paper incorporates various human learning strategies to further accommodate human learners.

We first employ *scaffolding* from constructivist learning theory to encourage demonstrations that are not just informative but also comprehensible. Specifically, we assume a general human learner without prior knowledge and sequence demonstrations that incrementally increase in informativeness and difficulty. Noting the cognitive science literature that suggests humans favor simple explanations that follow a discernible pattern, we also optimize for visual *simplicity and pattern discovery* when selecting demonstrations. Finally, toward effective *testing* of the learner's understanding, we show that the measure of a demonstration's informativeness during teaching can be inverted into a measure of expected difficulty for a human to predict that exact demonstration during testing.

A user study yields favorable results for the optimization of simplicity and pattern discovery, and for our measure of test difficulty. However, it yields mixed results for our approach of scaffolded teaching and provides insights regarding its shortcomings. Thus, we end with a reflection on how to further improve the proposed methods in future work.

2 Human Teaching Techniques

We take inspiration from constructivist learning theory and cognitive science in informing how a robot may teach a skill to a human learner so that the learner may correctly reproduce that skill in new situations.

Scaffolding: Scaffolding is a well-established pedagogical technique in which a more knowledgeable teacher assists a learner in accomplishing a task currently beyond the learner's abilities, e.g. by reducing the degrees of freedom of the problem and/or by demonstrating partial solutions to the task [7]. Noting the benefits seen by automated scaffolding to date (e.g. [8]), we implement the first of Reiser's [9] recommendations for software-based scaffolding, which is to reduce the complexity of the learning problem through additional structure. Specifically, we incorporate this technique when teaching a skill by providing demonstrations that sequentially increase in complexity and informativeness.

Simplicity and pattern discovery: Studies on explanations preferred by humans indicate a bias toward those that are simpler and have fewer causes [10]. Furthermore, Williams et al. [11] found that explanations can be detrimental if they do not encourage the learner to discover patterns or even mislead them with false patterns. We thus also optimize for simplicity and pattern discovery when selecting demonstrations that naturally 'explain' the underlying skill.

Testing: Effective scaffolding requires an accurate diagnosis of the learner's current abilities to provide the appropriate level of assistance throughout the teaching process [12]. A common diagnostic method is presenting the learner with tests of varying difficulties and assessing their understanding of a skill. Toward this, we propose a way to quantify the difficulty of a test that specifically assesses the student's ability to predict the right behavior in a new situation.

3 Implementation of Human Teaching Techniques

The robot's environment is represented as a Markov decision process (MDP). The robot has an optimal policy π^* (i.e skill) that maps each state s to the action that will optimize the sum of rewards R over an infinite horizon. With no loss of generality[1], we assume that R can be expressed as a weighted linear combination of reward features extracted from the states, i.e. $R = \mathbf{w}^\top \phi(s)$ [13]. For a handwriting policy, ϕ could include the completeness of letters, spacing between letters, etc. We assume that the learner is aware of ϕ but not $\mathbf{w}^*$ (nor how the tradeoffs in $\mathbf{w}^*$ manifest in behavior). Thus, how clearly a demonstration reveals $\mathbf{w}^*$ naturally correlates with the demonstration's informativeness.

Brown et al. [6] show that using IRL [14], demonstrations of a policy may be translated into half-space constraints that constrain the set of reward weights for which the policy is still optimal (i.e the behavioral equivalence class, or BEC). Assuming that conveying $\mathbf{w}^*$ is a good proxy for conveying the optimal policy, the BEC area can be used to quantify a demonstration's informativeness.

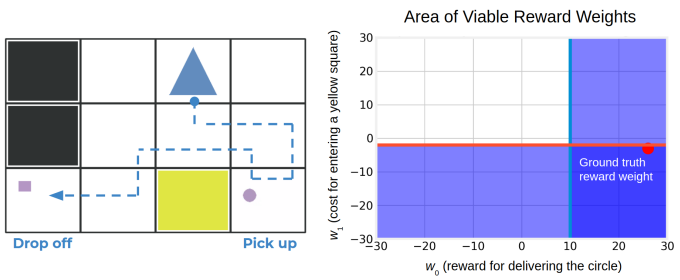

Figure 1: A demonstration of an optimal policy (left) can be translated into a set of half-space constraints on the possible underlying reward weights (right) [6].

Consider an example in the delivery domain in which the following weights $\mathbf{w}^* = [26, -3, -1]$ are assigned to the following reward features ϕ: delivering the package, entering mud, and taking an action. The demonstration in the left image of Fig. 1 corresponds to the constraints in the right image. With a unit cost for each action, the constraints on viable reward weights intuitively indicate that 1) $w_0^* \geq 10$ since a total of 10 actions were taken in the demonstration and that 2) $w_1^* \leq -2$ as the detour around the mud took two actions. Though this demonstration is fairly informative, another is needed to lowerbound the cost of the mud (for there are instances when the robot will go through mud than take an extended detour), and so on.

Scaffolding: Brown et al. [6] proposed the Set Cover Optimal Teaching (SCOT) algorithm for selecting the minimum number of demonstrations that results in the smallest BEC area. A pure IRL learner is assumed to fully grasp these few highly nuanced examples that delicately straddle decision-making boundaries and find any other demonstrations redundant. However, we posit that the BEC area of a demonstration not only inversely corresponds to the amount of

[1]These assumptions are rather generous, as any autonomous agent (embodied or not) in any location can in theory express its environment and its skill like so and utilize the methods outlined in this section.

information it contains about the possible values of $\mathbf{w}^*$, but also inversely corresponds to the effort required for a human to extract that information. Thus humans will likely benefit from additional scaffolded examples that ease them in and incrementally relax the degrees of freedom of the learning problem.

We begin scaffolding by first sorting all possible demonstrations in order from those with the smallest BEC area (most informative but hardest to understand) to those with the largest BEC area (least informative but easiest to understand). Then the demonstrations are clustered (we used k-means, with k=6), so that one may show demonstrations from every other cluster (to ensure that consecutive demonstrations have sufficiently different areas) from the largest to smallest BEC area. Finally, demonstrations selected by SCOT are shown to convey the maximum information.

Simplicity and pattern discovery: Though the BEC area of a demonstration provides an unbiased, quantitative measure of the information transferred to a pure IRL learner, humans are likely also influenced by the medium of the demonstration, e.g. the visuals. Visible differences between sequential demonstrations can highlight critical features, while visual clutter that does not influence the robot's behavior (e.g. extraneous mud not in the robot's shortest path) may distract or even mislead the human.

We first encourage pattern discovery by favoring a new demonstration that is visually similar to the previous demonstration. The aim is to highlight a change in environment (e.g. a new mud patch) that caused the change in behavior (e.g. robot takes a detour) while keeping all other elements constant. Finally, a measure of visual simplicity is also manually defined for each domain (e.g. the number of mud patches in a delivery MDP) and out of candidate demonstrations that are equally similar, the visually simplest is selected.

Testing: A demonstration's BEC area intuitively captures its informativeness during teaching; the smaller the area, the less uncertainty there is regarding $\mathbf{w}^*$. We propose a complementary and novel idea that the BEC area can be inverted as a measure of a demonstration's *difficulty as a test*, i.e. when a human is asked to predict this optimal behavior in a new environment. Intuitively, a large BEC area indicates many viable reward weights for a demonstration, and thus the human does not need to precisely understand $\mathbf{w}^*$ to correctly predict the optimal behavior. We can also use this measure to scaffold tests of varying difficulties to gauge the human's understanding of $\mathbf{w}^*$ and subsequently π^*.

4 User Studies

We ran an online user study that involved participants learning how to play three gridworld games (i.e. three domains), where each had a unique optimal strategy based on the rewards and costs associated with various objects and locations. For each game, the participants would first watch the robot teacher provide a few demonstrations of optimal gameplay. Then they were asked to play a few unseen instances of the game of varying difficulties, being assessed on whether or not they correctly reproduced the optimal behavior. For each test, participants were also asked to rank their confidence in their response on a 5-point Likert scale. Thus the study had two within-subject variables: *domain* and *test difficulty* (low, medium, and high, determined by the BEC area of the test).

The study primarily explored how incorporating human learning strategies when selecting demonstrations impacts a human's ability to correctly reproduce the optimal policy. Specifically, it examined how the presence and direction of scaffolding, and optimization of visuals, would impact the human's test performance. The between-subjects variables were *scaffolding class* (none, forward, and backward), and *visual optimization* (positive and negative, corresponding to the maximization and minimization of both simplicity and pattern discovery, respectively). For scaffolding class, forward scaffolding showed low, medium, then high information demonstrations (from the 5th, 3rd, and 1st BEC clusters), backward scaffolding showed forward scaffolding's demonstrations in reverse, and no scaffolding showed all high information examples from the 1st BEC cluster. A total of five demonstrations were shown for each domain, always ending with demonstrations determined by SCOT.

5 Results

108 participants were recruited using Prolific [15]. Participants self-reported age (18 to 52, M = 26.57, SD = 8.33) and gender (roughly 64% male, 34% female, 2% non-binary). Each of the six possible between-subjects conditions studies were randomly assigned 18 participants, and the order of the domains presented to each participant was counterbalanced. The domains were designed to be representative of a range of difficulties, and we subsequently average each participant's test scores across the domains in all following analyses.

Scaffolding: A two-way mixed ANOVA on test performance revealed a significant interaction effect between scaffolding class and test difficulty ($F(4, 210) = 2.79, p = .03$). Post-hoc pairwise Tukey analyses showed that no

scaffolding yielded significantly better test performance than forward scaffolding for high difficulty tests ($p = .05$, Fig. 2). Though not statistically significant, a trend of forward and backward scaffolding outperforming no scaffolding on low and medium difficulty tests was observed as well. Surprisingly, scaffolding class did not have a significant effect on overall test performance ($F(2, 105) = 0.02, p = .98$).

Simplicity and pattern discovery: A two-way mixed ANOVA on test performance revealed a significant interaction effect between visual optimization and test difficulty ($F(2, 212 = 5.61, p = .004$). Exploring the interaction effect with Tukey analyses revealed that optimized visuals had no effect on test performance on low ($p = .24$) and medium ($p = .90$) difficulty tests, but led to a significant improvement in performance in high ($p = .001$) difficulty tests for positive visual optimization ($M = 0.45$) over negative ($M = 0.31$).

Testing: A one-way repeated measures ANOVA revealed a statistically significant difference in a human's performance across test difficulty ($F(2, 214) = 162.22, p < .001$). Tukey analyses further revealed significant differences between each of the three groups, with the percentage of correct responses dropping from low ($M = 0.88$), to medium ($M = 0.66$), to high ($M = 0.38$) test difficulties ($p = .001$ in all cases). Spearman's rank-order correlation further showed a significant inverse correlation between test difficulty and participants' confidence in their responses ($r_s = -.39, p < .001, N = 324$).

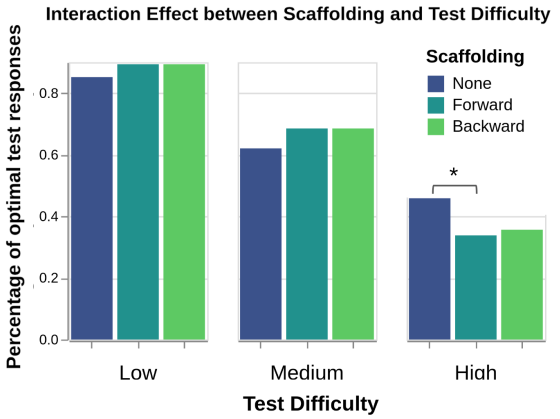

Figure 2: The influence of scaffolding on learner test performance depends on test difficulty.

6 Discussion and Future Work

Scaffolding: The unexpected result of backward scaffolding's on-par performance with forward scaffolding suggests that though BEC area is a theoretically well-motivated model for a demonstration's informativeness to a human, it is insufficient and our scaffolding order likely was not clear cut in either direction. Imagine providing the same example twice to a human; BEC area would incorrectly conclude that the second example was just as informative as the first.

Furthermore, forward and backward scaffolding (each comprised of low, medium, and high information demonstrations) yielding higher performance for low and medium difficulty tests over no scaffolding (comprised of only high information demonstrations), and the relationship being reversed for high difficulty tests, suggests that model-free, imitation learning-type (IL) reasoning may have also been at play. Though IL is another well-accepted model of how humans learn from demonstrations or experience (see [16, 17]), it is not obvious when when humans will use IL or IRL. Considering the observation made by Lage et al. [17] that people may be more inclined to use IL than IRL in less familiar situations, our participants may have benefited from more extensive pre-study practice and/or additional informative demonstrations.

Future work: We propose two directions for future work. First, we note that our selected demonstrations often revealed information about multiple reward weights at once, which could be difficult to decipher. Instead, we can further scaffold by teaching about one weight at a time, when possible. Second, Reiser [9] suggests that scaffolding should not only provide structure that reduces problem complexity but at times induce cognitive conflict to challenge and engage the learner. The current method of scaffolded teaching assumes that the learner has no prior knowledge. However, we observed that a subsequent demonstration with smaller a BEC area was sometimes easier to understand because the robot behaved exactly according to the model we had naturally developed from observing previous demonstrations. We believe that providing demonstrations which incrementally deviate from the human's current model will be more informative to a human and would be better suited to scaffolding.

Simplicity and pattern discovery: Optimizing visuals improved test performance, but only for high information demonstrations. Perhaps simplicity and pattern discovery could only produce a meaningful reduction in complexity for high information demonstrations, while those of low and medium information were already comprehensible.

Future work: Interestingly, memory arose as an unforeseen confounder resulting from the study procedure. As participants could not rewatch previous demonstrations (to enforce scaffolding order), consecutive demonstrations optimized for visual similarity sometimes led to greater confusion as participants believed they saw different behaviors in the same environment. Future iterations would benefit from 'markings of critical features' as suggested by Wood et al. [7], e.g. by explicitly highlighting the differences between consecutive environments with visual markers.

Testing: Objective and subjective results strongly support BEC area as a measure of test difficulty for human learners.

Future work: As previously noted, effective scaffolding is contingent on maintaining an accurate model of the learner's current abilities. Though this work assumed disjoint teaching and testing phases, learning is far more dynamic in reality. Important topics to explore thus include selecting an initial set of tests that can accurately discern the learner's current knowledge, and knowing when to switch between teaching and testing throughout the learning process.

7 Code and data availability

The code for generating demonstrations using the aforementioned methods can be found in the following repository: https://github.com/SUCCESS-MURI/machine-teaching-human-IRL.

The code for generating the user study (including videos of the teaching and testing demonstrations) and the data corresponding to our results can be found in the following repository: https://github.com/SUCCESS-MURI/psiturk-machine-teaching.

8 Acknowledgements

This work was supported by the Office of Naval Research award N00014-18-1-2503 and Defense Advanced Research Projects Agency (DARPA) / Army Research Office (ARO) award W911NF-20-1-0006. The views and conclusions contained in this document are of the authors and should not be interpreted as representing official policies, expressed or implied, of DARPA, ARO, or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

Finally, we would like to thank Vignesh Rajmohan and Meghna Behari for their assistance in creating the user study, and Pallavi Koppol for sharing her user study and data analysis templates.

References

[1] Anara Sandygulova, Wafa Johal, Zhanel Zhexenova, Bolat Tleubayev, Aida Zhanatkyzy, Aizada Turarova, Zhansaule Telisheva, Anna CohenMiller, Thibault Asselborn, and Pierre Dillenbourg. Cowriting kazakh: learning a new script with a robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 113–120, 2020.

[2] Arzu Guneysu Ozgur, Ayberk Özgür, Thibault Asselborn, Wafa Johal, Elmira Yadollahi, Barbara Bruno, Melissa Skweres, and Pierre Dillenbourg. Iterative design and evaluation of a tangible robot-assisted handwriting activity for special education. *Frontiers in Robotics and AI*, 7:29, 2020.

[3] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

[4] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

[5] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[6] Daniel S Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7749–7758, 2019.

[7] David Wood, Jerome S Bruner, and Gail Ross. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100, 1976.

[8] Sandra Sampayo-Vargas, Chris J Cope, Zhen He, and Graeme J Byrne. The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, 69:452–462, 2013.

[9] Brian J Reiser. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences*, 13(3):273–304, 2004.

[10] Tania Lombrozo. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759, 2016.

[11] Joseph Jay Williams, Tania Lombrozo, and Bob Rehder. Why does explaining help learning? insight from an explanation impairment effect. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[12] Allan Collins, John Seely Brown, and Susan E Newman. Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. *Thinking: The Journal of Philosophy for Children*, 8(1):2–10, 1988.

[13] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[14] Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *in Proc. 17th International Conf. on Machine Learning*, 2000.

[15] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[16] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

[17] Isaac Lage, Daphna Lifschitz, Finale Doshi-Velez, and Ofra Amir. Exploring computational user models for agent policy summarization. *arXiv preprint arXiv:1905.13271*, 2019.