Policy Shaping in Domains with Multiple Optimal Policies (Extended Abstract)

Himanshu Sahni [†]
himanshu@gatech.edu

Brent Harrison [†]
brent.harrison@cc.gatech.edu

Kaushik Subramanian [†]
ksubrama@cc.gatech.edu

Thomas Cederborg [†]
thomascederborgsemail@gmail.com

Charles Isbell [†]
isbell@cc.gatech.edu

Andrea Thomaz [§]
athomaz@ece.utexas.edu

ABSTRACT

In many domains, there exist multiple ways for an agent to achieve optimal performance. Feedback may be provided along one or more of them to aid learning. In this work, we investigate whether humans have a preference towards providing feedback along one optimal policy over the other in two gridworld domains. We find that for the domain with significant risk to exploration, 60% of our participants prefer to discourage the agent's exploration along the risky portion of the state space, while 40% state that they have no preference. We also use the interactive reinforcement learning algorithm Policy Shaping to evaluate the performance of simulated oracles with a number of feedback strategies. We find that certain domain traits, such as risk during exploration and number of optimal policies play an important role in determining the best performing feedback strategy.

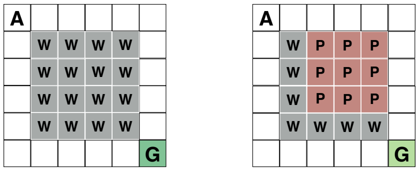
Keywords

interactive machine learning; learning from critique; reinforcement learning; policy shaping

1. INTRODUCTION

In real world environments, an expert may not always be at hand to train the behavior of a human guided learning agent. Thus, such agents must be able to learn from feedback by non-expert teachers. Policy Shaping [2] attempts to use feedback directly as policy advise and combine it with standard reinforcement learning approaches such as Bayesian Q-learning. Recently, Policy Shaping has been shown to be robust to noisy and sparse human feedback [1]. Cederborg et al. discovered that human provided critique led to better performance than the simulated teacher on two different pac-man domains. They state that the difference in performance could be because humans seemed to give positive feedback to any strategy that appeared optimal. The simulated teacher, on the other hand, had a single fixed optimal policy that was computed beforehand. It may also be the case that certain optimal policies can be learnt faster and the human teachers were guiding the agent towards them. Their experimental setup was not designed to verify these conjectures, since the finding came as a surprise. In this work, we specifically investigate whether humans employ differing feedback strategies in two different gridworld domains. Knowledge of whether different people chose different optimal policies to teach in the same domain can be used to lend us greater insight into designing algorithms learning from human critique. We conduct a user study in which participants are asked to provide feedback to an agent acting on a pre-determined set of policies. In a post-study interview, we ask the users which, if any, optimal policy they gave preference to during teaching. We find that users do have feedback preferences in one of the domains.

The findings of the user study motivate an in-depth analysis of how learning performance is impacted by choice of feedback strategies over multiple optimal policies. A complicated task in a real-world environment will likely have many ways of solving it. A non-expert, end user critiquing the agent may provide feedback for all of them at the same time. Or they may choose to ignore all but one in order to simplify their task of teaching the agent. A human teacher may also try to adapt their feedback to encourage more of what the agent is already doing correctly. Our hypothesis is that different tasks will have different optimal feedback strategies. Therefore, an understanding of which strategies lead to best performance in which domains is important.



(a) *blocks* domain (b) *pits* domain

Figure 1: The gridworld domains used in user studies and simulated experiments. The *blocks* domain has two identical optimal policies while in *pits* one of them is clearly better for learning in combination with a random exploration strategy.

2. RELATED WORK

Interactive machine learning is emerging as a promising field with many useful applications. In the context of reinforcement learning, prior work has attempted to convert feedback signals into rewards, akin to those coming from the environment. However, human feedback may be noisy, inconsistent, and not easily translated into rewards. An alternative solution has been to use feedback to influence the policy, instead of the reward signal. Policy Shaping [2] ex-

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

[†] College of Computing, Georgia Institute of Technology

[§] Department of Electrical and Computer Engineering University of Texas at Austin tracts policy level information directly from feedback and combines it with environment based signals.

This work explores the effects of using Policy Shaping in situations where there are multiple optimal ways of behaving. A modification to Q-learning, $\tilde{Q}$-learning, performs optimally when combined with an exploration strategy [3]. John notices that paths along the wall of a gridworld can be faster to learn as a sub-optimal move has less of a chance of sending the agent further away from the goal. Our work explores a similar effect in the setting of learning from critique. We try to answer the question: when there are multiple solutions to a task, all equally optimal, what effect does feedback strategy have over learning performance?

3. HUMAN FEEDBACK PREFERENCE

The goal of our user study was to determine whether people have differing preferences over feedback strategies when faced with multiple optimal policies in the same domain. We recruited 10 participants, all graduate students not familiar with this project, who were presented with an agent acting on the same pre-determined policies in the *pits* and the *blocks* domains (Figure 1). The participants were instructed to provide positive or negative critique (Figure 2).

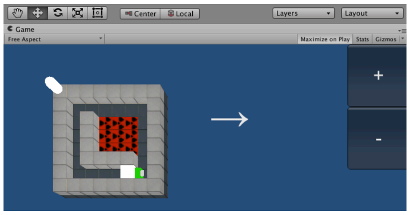


Figure 2: Feedback interface. The agent is represented by the white, the goal by the green and the pits by the red blocks. An arrow indicates the most recent action.

At the end of the study, the subjects are asked a small set of questions about their preferred feedback strategy. In the *pits* domain 6 of our participants said that they preferred the safer path to the goal. The most common reason was that it was "less dangerous" or it had "less chance of [the agent] falling into the pit". The remaining 4 said that they had no preference over the path the agent took. They said, "[I] wanted the agent to reach the goal as fast as possible" or "feedback I was giving was based on its [agent's] first action". Participant 6 said, "I wanted the agent to know more than one path". All participants said that they had no preference between the paths in the *blocks* domain and that "they looked identical". Since users may chose different feedback strategies depending on the domain, this leads us to the question of how this affects learning performance. We explore this using simulated oracles in the next section.

4. ORACLE PERFORMANCE

We study a carefully designed set of teachers that span a range of evaluation behaviors. Automated oracles provide feedback consistent with each behavior and allow us fine control over its frequency. Broadly speaking, we consider four different kinds of feedback strategies. The *all policy* oracle provides positive critique along all optimal policies while the *single policy* oracle does only along one. The *single path* oracle provides positive critique along only one optimal trajectory through the state space. In states not on the trajectory, it does not provide any feedback at all. The *adaptive* oracle is more likely to positively critique optimal actions that the agent has tried before.

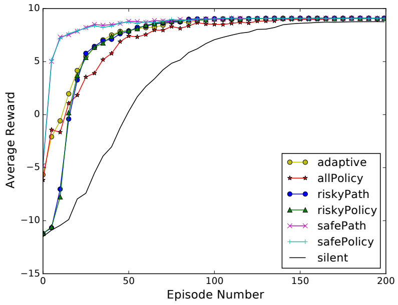


Figure 3: Average rewards (1000 independent trials) obtained by each oracle in the *pits* domain. T-tests are used to ensure statistical significance. *Risky policy* and *risky path* behave as *single path* oracles on the path along the pits.

Figure 3 shows the performance of the oracles and when no feedback is provided (*silent*) in the *pits* domain. As expected, the *safe path* and *safe policy* oracles converge in the least number of episodes. Interestingly, the *risky path* and *risky policy* oracles also converge faster than the *all policy* oracle. A one-sample t-test rejects the null hypothesis that the average reward achieved by *risky policy* oracle is optimal until episode 109 ($p < 0.05$). For *all policy* oracle this occurs at episode 177. The *risky policy* oracle is curtailing the agent's exploration and it learns the *risky policy* faster than it manages to learn a policy over the entire state space. In a larger version of the *blocks* domain with wider corridors, the *single path* oracle is slowest to converge ($p < 0.05$). This may be due to the large number of optimal policies of which only a single path is being critiqued.

5. CONCLUSIONS AND FUTURE WORK

In the *pits* domain, four of our users stated that they have no preference over the two paths. With simulated oracles, we have shown that in fact the *all policy* feedback strategy is slower to achieve optimal reward than any of the *single path* ones. In general, it may be the case that the policy which is easiest to learn is not the one the teacher is interested in providing feedback on. A lot of computational and human effort may be wasted in exploring difficult to learn optimal policies before positive results are seen. In depth exploration of best performing teaching strategies in different domains is a fruitful and important direction of future research.

Acknowledgments

This work was funded under ONR grant number N000141410003

REFERENCES

[1] T. Cederborg, I. Grover, C. L. Isbell, and A. L. Thomaz. Policy Shaping With Human Teachers. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[2] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. L. Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems 26*, pages 2625–2633. Curran Associates, Inc., 2013.

[3] G. H. John. When the best move isn't optimal: Q-learning with exploration. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, page 1464, 1994.