Retro-TAMER: Retroactive Feedback Assignment within the TAMER Framework

Andrew Wang, Gian Marco Visani

1 Introduction

TAMER (Training an Agent Manually via Evaluative Reinforcement [4]) is a framework that enables a human trainer to interactively shape an agent's policy via reinforcement signals. The human observes the agent as it is interacting with the environment, and, at any given time, they can choose to give either positive feedback or negative feedback, based on whether they want to encourage or discourage certain actions that the agent has performed. The main goal of TAMER is to reduce the sample complexity for learning a "good" policy, and doing so via a simple and intuitive shaping mechanism that allows lay users to give useful feedback regardless of technical background. We believe the most important contributions of TAMER are the following: its accessibility, as mentioned above, and its ability to push the agent in the right direction at the early stages of training, thus reducing the cost of the initial learning trials, which would otherwise be performing random actions.

We believe the TAMER framework is very promising, and for this very reason we would like to contribute to its development by addressing what we believe are potential areas of improvement. First, we believe that it is too hard for a human to *reliably* evaluate the performance of an agent *online*. In particular, it is hard for a human to quickly evaluate series of actions, or at least to do it so as quickly as they are required by the current TAMER framework. Series of actions are often more important than individual actions in achieving a goal, and thus properly assign feedback to them can significantly boost learning. This lack of reliability, or noisiness, of feedback might be what is limiting the peak performance of the current TAMER agents [4].

Second, we believe that the confidence on the feedback given needs to be put into consideration, or at least explored.

We first attempt to find a solution to the feedback's noisiness by allowing the human to assign feedback retroactively, i.e. after a fixed time frame. The human is allowed to review TAMER's interaction with the environment within the time frame on a recording, and assign feedback as specific times using a scroll bar. This way, the human can better review the behavior of the agent at their own pace, and also better estimate the impact of series of actions, which they can reward by assigning positive impact at the very end of them, or throughout them. We believe this approach would make the human feedback much more

reliable. However, we are also aware of its shortcomings. In particular, in this approach TAMER does not learn on the fly, i.e. its policy does not improve within the time frame, as it would be expected to do with regular TAMER. With our approach, TAMER does not make increasingly better actions, but rather it improves its policy in larger steps. We believe that learning might be slower if done this way, and thus that the length of the time frame needs to be fine tuned in order to make the benefits of increased reliability worth-wile. In section 3, we will describe the experiments we plan to perform in order to fine-tune our approach.

2 Background Related Work

Following, you should provide the necessary background and discuss related work in the RL literature. This section should also be about a page. Citations should be in BibTeX format [**?**].

TAMER is introduced prominently in Knox and Stone [4]. As mentioned in the introduction, as a framework it allows humans to advise a reinforcement learning agent on its choice of actions. This allows human participants who are undoubtedly more initially familiar with the problem an RL agent is tasked with solving to push the agent in the right direction. TAMER specifically deals with binary positive/negative feedback labels, and this makes TAMER very easy to use for lay people that do not have to be intimately familiar with reinforcement learning as well. In addition, not only does TAMER allow agents to operate in a reward sparse environment, providing a sort of jump-start to reward states, but eliminates the need for an environmental reward function in the first place, and allows for learning to be done in complex environments that would take autonomous agents too much time and computational resources to learn effectively in.

We are particularly interested in the performance of Tetris by the TAMER agents in Knox and Stone; notably, although TAMER is able to learn very quickly how to clear lines (under three games), its peak performance pales in comparison to methods such as genetic algorithms, which on average clear nearly 600,000 lines as compared to TAMER's 65.89 lines. In part, our motivation for continuing to look for improvements on top of TAMER is a following paper by Knox and Stone [5]. In this paper, TAMER is truly used in conjunction with eight other reinforcement learning models to demonstrate that TAMER+RL methods outperform SARSA(λ) in both cumulative reward and peak performance, even if it's not done in the context of Tetris specifically. However, this gives us confidence that having TAMER result in a higher peak performance would potentially allow for better end results when TAMER is combined with other reinforcement learning methods, to be explored in future research.

Our concentration on TAMER in particular is further motivated by [1], which showcases the power of TAMER in being able to teach arbitrary tasks absent of inherent environmental reward. Similarly to Knox and Stone[5] as well, [3] shows an addition of learning feedback for intended actions for TAMER agents,

showing further existence of available continuation with TAMER as a methodology.
We also considered other methods of human-based policy shaping, such as the Advise method demonstrated in Griffith et. al[2], which while also allow for human-in-the-loop RL, have problems such as many hyperparameters to tune, being too domain specific, or having to built off of existing RL algorithms. TAMER in the end was more attractive because of its strengths, as well as the existence of a lot of literature discussing potential improvement on top of TAMER.

3 Technical Approach / Methodology / Theoretical Framework

We will implement TAMER as it is described in [4]:

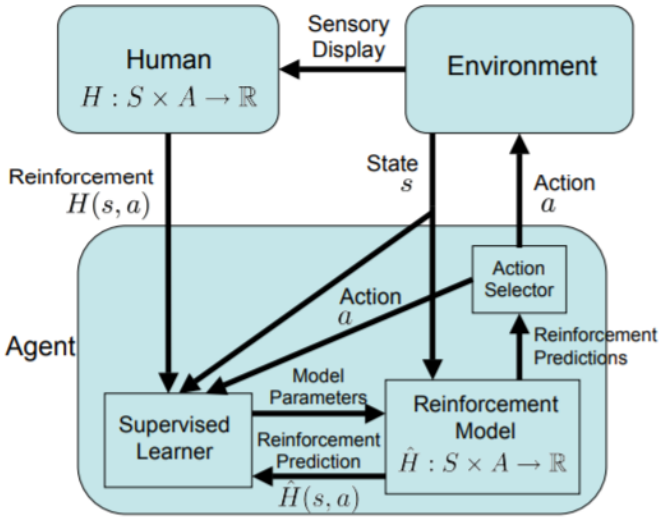

Figure 1: Framework for Training an Agent Manually via Evaluative Reinforcement (TAMER).

Figure 1: TAMER Framework Methodology

Knox and Stone do not specify which supervised model is used. We plan on asking advice to Jivko or directly ask the authors themselves which model they used. Otherwise, we will try different models and aim at replicating the results in [4]. We expect the models to be non-linear, so two viable models are a neural network (harder to fine-tune) and a random forest (potentially less powerful). We will test mainly on the Tetris domain, as it is done in [4], so that we can make a comparison with the results.

3.1 Experiments

We plan on performing three main experiments:

1. Normal TAMER, as a baseline.
2. Retroactively assign feedback by having the human look back at the agent's interaction with the environment within a specific time frame. We will need to fine-tune the length of the time frame.
3. Run normal TAMER for as long as a whole episode, and then have the human look back at the agent's actions and assign feedback again. This allows a fast learning that can be further reinforced, corrected, or, more generally, fine tuned later on.

We are considering fine-tuning the length of the time frame for experiment 2 ourselves for the sake of time efficiency. Then, we plan on gathering some people and have them perform all three experiments, in order to gather our data consistently. Furthermore, we will have them play the game of Tetris a bit by themselves, so that they can get accustomed to it beforehand and not learn how to play it as they are performing the experiments. We will also change the order in which participants do the three experiments, to furthermore ensure that results aren't being skewed towards the last experiment being performed.
At the end of each experiment, we will ask the participants to rate, on a scale from 1 to 10, their level of confidence on the feedback they provided during the experiment.

3.2 Challenges

We believe the greatest challenge will be in mapping the moment in which the participant delivers feedback, as they are seeing it in the replay, to the particular action they were targeting as it is represented in the algorithm. We will need to store the whole history of agent-environment interactions within the time frame, and allow the scroll bar to search along such history.
Another challenge would be consistently comparing results of the three models we have: we want to make sure that if we use the same participant over the course of the three experiments that they are not improving as players and thus skewing results towards the last experiment. We attempt to address this by having players familiarize themselves with the game beforehand and change the order of the experiments they perform, but we will still need to take this potential issue into consideration when evaluating our results.
One final challenge Would be to make sure our initial results for TAMER are consistent with the experimental results from [4]. This is not strictly necessary, and it's very possible that due to slightly different implementation of TAMER

or the performance of the participants themselves that we will not be able to match up our initial results with that of the paper, but it will make our results all the more convincing and significant if we can match the baseline set by Knox and Stone, and then show either improvement or disimprovement by the methodologies we are exploring.

4 Evaluation

Again, the goal of our experiments is to improve upon the results the of the TAMER agent on the game of Tetris, as described in Knox and Stone [4]. The specific metrics of interest are the performance at peak (so in this environment, how many lines are cleared in the best observed episode), in addition to how much time (in both number of total episodes as well as total time steps across all of the episodes) it takes to reach the peak performance. We will also be interested in how long it takes for our agents to reach the peak performance of the agent described in Knox and Stone [4], or if they do at all.
The success of our experiments will be determined in two ways:

1. First, we would be interested in establishing a baseline for our experiments, and this will be done by comparing our agents to the TAMER agent described in Knox and Stone by the metrics they describe in their table of results, namely: Mean lines cleared at game 3 (65.89), Mean lines cleared at peak (65.89), and Games for peak (3). If our baseline TAMER model can achieve comparable performance results to these, then our final results for the other two experiments will be immediately significant so long as they are implemented correctly, regardless of whether they perform better or worse.

2. Second, we will be comparing the results of our second and third experiments to the results of our own baseline TAMER agent. For this comparison, it does not matter as much whether or not our baseline TAMER agent matched up appropriately against the agent in the paper; we are mostly interested in the relative performances of the three methodologies. In addition to populating a table with the three metrics from Knox and Stone described above, we will be looking at a number of more nuanced metrics to get a more valuable and meaningful comparison between these three methods which have directly corresponding aspects of their implementation. As per our goals, we will be plotting:

    - Average cumulative lines cleared over number of games;
    - Average cumulative lines cleared over number of individual time steps ("choices" made by the agent, both during playing time and replay by human participants); and
    - If time/feasibility permits: average cumulative lines cleared over real human time.

We predict that our second experiment, having users only assign feedback to choices made by the agent retroactively will improve the peak performance of the agent in terms of average lines cleared without much change in the total number of games it takes to reach that peak performance. However, we expect average cumulative lines cleared over time steps will show poorer performance as each game will essentially have to be played over a second time to reach the equivalent amount of training as the baseline, so graphs of average cumulative lines cleared over individual time steps should not only appear more step-like due to not assigning reward on the spot, but also have a less steep slope on average. Of course, this will take much longer than the baseline in terms of real human time, due to human participants going back and replaying the episode, multiple times if need be, to determine the most accurate advice to give to the agent.

In our third experiment, with having users assign feedback in line with the implementation of TAMER, but also having users go back and retroactively adjust their feedback to the agent, we again expect both peak performance and games to peak performance to be higher than the baseline, but about on par with the second experiment. Average cumulative lines cleared should be better over all time metrics than the second experiment because the agent is both actively learning and applying its newfound knowledge at each time step in addition to having additional human feedback given retroactively to improve the quality of the initial feedback given. Because of this, average cumulative lines cleared over time steps is predicted to even surpass that of the baseline. In terms of real human time though this method will still perform worse than the baseline, because the human participants have to go back through a second time to reevaluate their initial feedback. We do predict that it will take less real human time for the third experiment than the second though because the human participants will be evaluating situations they are already familiar with and reviewing thought processes that should be relatively fresh in their minds, rather than having to create from scratch an evaluation of the agent's decision.

Our experimental methodologies will be proven to be a significant performance over regular TAMER as described in Knox and Stone so long as the peak performance improves over the baseline in experiments two and three without sacrificing number of games to reach that peak performance, and critically, if experiment three is able to beat both experiment two and the baseline critically in terms of average cumulative lines cleared over time steps, that will definitively show the value of reevaluating human feedback retroactively as the cost of improving performance is shown to not be hindered too much in terms of time. The final metric to potentially consider, real human time, will end up being a more subjective metric to weigh performance by, and since we are not certain on the scale of how much longer retroactive feedback will take as well as potential issues such as UI-friendliness having a large impact on human time taken to retroactively give feedback, we will plan on developing a more refined way to evaluate the impact of real human time as a metric as we start to perform the experiments.

5 Timeline and Individual Responsibilities

Within one week, we hope to settle on the environment we will use for the project. We are currently considering either the OpenAI python environment for Tetris, or if possible we will ask Professor Jivko if we can get access to the Tetris environment used by Knox and Stone [4]. We will also get a Github repository set up within the week, and download any dependencies we will need to carry out the experiments.

Before the first checkpoint, we will code the baseline TAMER agent in the environment and test it ourselves to compare results to that of Knox and Stone's implementation. At this point we will also be outlining how we are planning to tackle the issue of retroactively assigning reward such that the human participants are sure to map the feedback they give to the choices the agent is making correctly.

Within the two weeks after the first checkpoint, we will have implemented a reliable way of mapping assigned feedback to agent choices, and we will have start setting up visuals corresponding to the states and actions in the game that the agent sees as well so the human participants will have a very easy time of seeing the exact game state and applying feedback appropriately.

Within the week after that, we will be testing ourselves our implementation of retroactive feedback, as well as combining retroactive feedback with a regular implementation of TAMER. During this time, we will also be figuring out at what intervals within a game that the game should be put on hold and the human participants can go back and provide feedback/reevaluate their feedback on the agent's actions.

We will take a week to gather participants to go through trials with the three experiments, and a final week to consolidate all of the trial data, generate the aforementioned graphs, tables, and figures, and complete the final write-up.

In terms of collaboration and individual responsibilities, the entire project appears to be a very "in series" process; there aren't many places where one of us can be working on one part while the other works on a different part because each section of the project relies on having complete background knowledge and experience with working on the previous part. Working with the environment and coding the baseline model require a good understanding of the background material as well as familiarity with the model; coding the second experiment will rely on familiarity with TAMER for the baseline model, and the third experiment will rely on knowing how to implement both the baseline TAMER model from the first experiment as well as adding in retroactive feedback from the second experiment. Sourcing participants for trials and having participants go through the experiments for gathering data will be an equal responsibility for the both of us, though we do not have to perform the trials at the same time

obviously. We will divide up the final data-consolidation, figure generation, and writing of the report in equal portions as well, as there's no particular reason to dictate one part of that process to any one person; this way we can ensure that we both understand the entire process and are able to articulate our results on our own if need be.

References

[1] Cynthia Breazeal. 69 tamer, Oct 2013.

[2] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2625–2633. Curran Associates, Inc., 2013.

[3] W Bradley Knox. Learning from feedback on actions past and intended. Citeseer.

[4] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM, 2009.

[5] W. Bradley Knox and Peter Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, pages 5–12, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.