Agent Robustness to Negative Interruptions: A Multi-agent Learning Approach

Mustafa Mert Çelikok
mustafa.celikok@aalto.fi

Tomi Peltola
tomi.peltola@aalto.fi

Samuel Kaski
samuel.kaski@aalto.fi
Helsinki Institute for Information Technology HIIT
Department of Computer Science, Aalto University, Helsinki, Finland

Abstract

Safety-critical autonomous systems often allow human users to interrupt their operation as a safe-guard measure. However, human decisions are influenced by cognitive biases and limitations which may lead to interruptions that result in sub-optimal outcomes. A robust agent must avoid actions which might trigger such interruptions. We formulate the problem of the negative interruptions as an instance of the asymmetric multi-agent reinforcement learning, where the user is able to observe the system's latest action and decide whether to interrupt its execution or not. Our formulation is closely related to robust reinforcement learning when interruptions are taken as disturbance. We propose a modified independent Q-learners solution where the user and the system learn their optimal policies together in single-agent perspectives of the interaction. Empirical results indicate that the convergence depends on the user's reward function, and opponent modelling is necessary for a general solution.

1 Introduction

Human decision-making deviates from rationality often due to cognitive biases and limitations. In order to explain these deviations, earlier works tried to propose alternatives to expected utility theory [1], and recently bounded rationality has emerged as a common framework for both machines and humans [2]. Autonomous learning systems promise to help humans make rational decisions devoid of negative cognitive biases. However most autonomous systems are not fully autonomous: they provide ways for humans to interrupt their operation.

Previous works have focused on developing reinforcement learning methods that are *safely interruptible* [3], where humans can interrupt the agent in order to avoid negative consequences. However, human interruptions are a part of the human decision-making and thus are affected by cognitive biases and limitations. In this work, we are interested in those interruptions that result in negative consequences. Negative interruptions are a form of the disuse of autonomous systems [4], where human operators may learn to ignore an autonomous system's recommendations or replace its actions with their own. Recently, c-intervention games have been proposed to model the negative interruptions where the system and the user disagree on the dynamics of the environment [5]. The user has an erroneous model of the dynamics, and can replace the system's actions as in interruptible policies. The proposed solution concept is the Stackelberg equilibrium which assumes: the system commits to a policy and, the user observes the system's entire future commitment (i.e. policy).

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NIPS), .

We re-formulate the negative interruptions problem as an instance of asymmetric multi-agent reinforcement learning [6], and make a complementary connection to robust reinforcement learning. Our main contributions are: **(1)** We constrain the user to observe only the system's latest action. After observation, the user may override the action with another. We believe our setting is more realistic for modelling human interruptions as it does not require full knowledge of the system's policy. The temporal intuition behind our choice is that the user and the system are operating in different time-scales which allows the user to observe and interrupt before the system completes its action. **(2)** We formulate the interruptions based on the disagreements between the agents' optimal Q functions instead of models, which allows us to apply model-free methods.

2 Problem Setting

Robust reinforcement learning aims at learning policies that are robust to bounded perturbations (also called disturbances) of the environment. The robust optimisation task is often modelled as a competitive game between an actor and a disturber [7]. In SA-rectangular robust RL [8], the actor chooses policy π that maximises its expected returns G under the disturbed dynamics $\bar{P}_{s,a} = \bar{P}(s'|s,a)$ with the reward function R, while the disturber tries to minimise the same quantity by disturbing the original dynamics of the environment $P_{s,a}$ into $\bar{P}_{s,a}$. The introduced disturbances are bounded for each (s,a) by $\epsilon_{s,a} \geq 0$ which results in the optimisation objective below;

$$\max_{\pi} \min_{\bar{P}} \{G(\pi, \bar{P}, R) : ||\bar{P}_{s,a} - P_{s,a}|| \leq \epsilon_{s,a}, \forall s, a\}.$$

Intuitively, this means the disturber can observe the last action and state (s,a), then choose a $\bar{P}_{s,a}$ since it is constrained for each (s,a) separately. Using this setting, we model the interrupting user as the disturber, and the interruptions as disturbances. Our disturber differs from its robust RL counter-part in two ways: (1) Instead of fully-competitive, it is a self-interested agent with its own decision process, (2) Instead of perturbing the transitions of the environment directly, it alters them by replacing the system's actions before they execute. We will refer to the user as the disturber and the system as the actor henceforth.

To see how the disturbances affect the transitions, define the single-agent problems of the actor and the disturber as their undisturbed MDPs $\mathcal{M}^{(A)} = (\mathcal{S}, \mathcal{A}^{(A)}, \mathcal{T}^{(A)}, \mathcal{R}^{(A)})$, and $\mathcal{M}^{(D)} = (\mathcal{S}, \mathcal{A}^{(D)}, \mathcal{T}^{(D)}, \mathcal{R}^{(D)})$ respectively, where $\mathcal{A}^{(D)} = \mathcal{A}^{(A)}$. Since the actor and the disturber are in the same environment, we will assume $\mathcal{T}^{(A)} = \mathcal{T}^{(D)} = \mathcal{T}$. Keep in mind that the actor and disturber do not have to be model-based, and $\mathcal{T}$ is simply where their transition observations would have come from if they were alone in the environment. The multi-agent interaction between the actor and the disturber is modelled by the Markov game $\mathcal{M} = (\mathcal{S}, \mathcal{A}^{(D)} \cup \{a^0\}, \mathcal{A}^{(A)}, \mathcal{T}', \mathcal{R}'^{(A)}, \mathcal{R}'^{(D)})$, where a^0 is called the null action, $\mathcal{T}'$ defines the joint-action transition dynamics, and $\mathcal{R}'^{(A)}, \mathcal{R}'^{(D)}$ are the modified reward functions of the actor and the disturber. If the disturber chooses any action other than null, it replaces the actor's action, otherwise the actor's action is executed [1]. The joint-action transitions and the modified reward functions are defined as;

$$\mathcal{T}'(s'|s, a^{(A)}, s^{(D)}) = \mathbb{1}(a^{(D)} = a^0)\mathcal{T}(s'|s, a^{(A)}) + \mathbb{1}(a^{(D)} \neq a^0)\mathcal{T}(s'|s, a^{(D)})$$
$$\mathcal{R}'^{(A)}(s, a^{(D)}, a^{(A)}) = \mathbb{1}(a^{(D)} = a^0)\mathcal{R}^{(A)}(s, a^{(A)}) + \mathbb{1}(a^{(D)} \neq a^0)\mathcal{R}^{(A)}(s, a^{(D)})$$
$$\mathcal{R}'^{(D)}(s, a^{(D)}, a^{(A)}) = \mathbb{1}(a^{(D)} = a^0)\mathcal{R}^{(D)}(s, a^{(A)}) + \mathbb{1}(a^{(D)} \neq a^0)[\mathcal{R}^{(D)}(s, a^{(D)}) - c(s)]$$

where $\mathbb{1}$ is the indicator function and $c(s)$ is the cost of interruption which effectively serves the same purpose as $\epsilon_{s,a}$[2]. If the disturber chooses the null action, actor's action gets executed and thus the disturber observes the reward $\mathcal{R}^{(D)}(s, a^{(A)})$. However, if the disturber chooses to interrupt with $a^{(D)} \neq a^0$, then it observes a penalised version of $\mathcal{R}^{(D)}(s, a^{(D)})$.

Even though a joint-action dynamics model, $\mathcal{T}'$ in essence models a dynamically perturbed single-action transition dynamics. If the actor commits to a policy $\pi^{(A)}$, the Markov game reduces to an MDP for the disturber. Similarly for the actor, if the disturber commits to a policy $\pi^{(D)}$, it reduces to a perturbed version of $\mathcal{M}^{(A)}$. If we assume that $\pi^{(D)}(a^0|s) = 0, \forall s$ then $\pi^{(A)}$ does not matter

[1] Null action allows us to subsume the interruption initiation function from [3] into the policy.

[2] Even though defined as $c(s)$ for notational convenience, it can be easily extended as $c(s,a)$.

since actor loses the control entirely. On the other hand, if $\pi^{(D)}(a^0|s) = 1, \forall s$ the actor's problem reduces to $\mathcal{M}^{(A)}$. We assume two limit behaviours with regard to the cost of interruption $c(s)$, $\lim_{c(s)\to\infty} \pi^{(D)}(a^0|s) = 1$ and $\lim_{c(s)\to 0} \pi^{(D)}(a^0|s) = 0$.

3 Algorithms

The dynamics of a multi-agent environment are non-stationary and non-Markovian from the individual perspectives of independent learners [9]. However, the transitions of the actor-disturber game $\mathcal{T}'$ have special properties. Let's assume the actor and the disturber are independent learners with $\mathcal{M}'^{(A)}$ and $\mathcal{M}'^{(D)}$ as their independent perspectives of the Markov game $\mathcal{M}$. The transitions of $\mathcal{M}'^{(D)}$ depend on the actor's action only for a^0, and follow $\mathcal{T}$ for any other action. The cost function $c(s, a^{(A)})$ effectively determines whether the disturber will choose a^0 at $(s, a^{(A)})$ or not. For the constant cost $c(s, a^{(A)}) = c$, a high enough c will make a^0 the optimal action for most states in $\mathcal{M}'^{(D)}$. In the end, the $\mathcal{M}'^{(D)}$ is non-stationary and non-Markovian only for the transitions that involve the null action, and c limits the non-stationarity of $\mathcal{M}'^{(A)}$.

We propose a modification to independent Q-learners where only the disturber's update is modified. If the disturber interrupts, it updates $Q^{(D)}(s, a^{(D)})$ with the penalised reward. However, if it takes the null action, it updates $Q^{(D)}(s, a^0)$ without the interruption penalty and performs a counter-factual update for $Q^{(D)}(s, a^{(A)})$, as if it had interrupted $(s, a^{(A)})$ with $a^{(A)}$. The counter-factual update discourages the disturber from learning to replace the actions with themselves. The general update rule is defined in algorithm 1. Different to the asymmetric Q-learning algorithm proposed in [6], our disturber does not maintain a copy of the actor's Q function.

Algorithm 1 Actor-Disturber Independent Q-Learning updates for $(s_t, a_t^{(A)}, a_t^{(D)}, s_{t+1})$

$Q^{(A)}(s_t, a_t^{(A)}) \leftarrow Q^{(A)}(s_t, a_t^{(A)}) + \alpha(\mathcal{R}'^{(A)}(s_t, a_t^{(D)}, a_t^{(A)}) + \gamma V^{(A)}(s_{t+1}) - Q^{(A)}(s_t, a_t^{(A)}))$
$Q^{(D)}(s_t, a_t^{(D)}) \leftarrow Q^{(D)}(s_t, a_t^{(D)}) + \alpha(\mathcal{R}'^{(D)}(s_t, a_t^{(D)}, a_t^{(A)}) + \gamma V^{(D)}(s_{t+1}) - Q^{(D)}(s_t, a^{(D)}))$
if $a_t^{(D)} = a^0$ **then**
  $Q^{(D)}(s_t, a_t^{(A)}) \leftarrow Q^{(D)}(s_t, a_t^{(A)}) + \alpha(\mathcal{R}'^{(D)}(s_t, a_t^{(A)}, a_t^{(A)}) + \gamma V^{(D)}(s_{t+1}) - Q^{(D)}(s_t, a_t^{(A)}))$
end if

4 Experiments

The underlying problem of our experimental task is cooperative where the actor is an interruptible autonomous system helping its user. The reward functions of the user is defined as $\mathcal{R}'^{(D)}(s, a^{(D)}, a^{(A)}) = \mathbb{1}(a^{(D)} = a^0)\mathcal{R}^{(A)}(s, a^{(A)}) + \mathbb{1}(a^{(D)} \neq a^0)[\mathcal{R}^{(A)}(s, a^{(D)}) - c(s)]$. For the ease of computation, we define a constant cost $c(s) = c$ for all the states. The discrepancy between the actor and the disturber is caused by the discounting. The actor uses exponential discounting, while the disturber uses hyperbolic. Hyperbolic discounting is known to capture time-inconsistency bias of human decision-makers [10] which leads to preference reversals. Different methods of discounting inevitably lead to different Q functions, and possibly to different

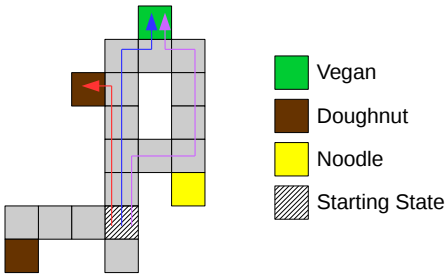

Figure 1: Single-agent optimal policies of the actor (blue) and the disturber (red) are different for the same starting state. The robust policy of the actor (purple) goes through the longer route to avoid interruptions.

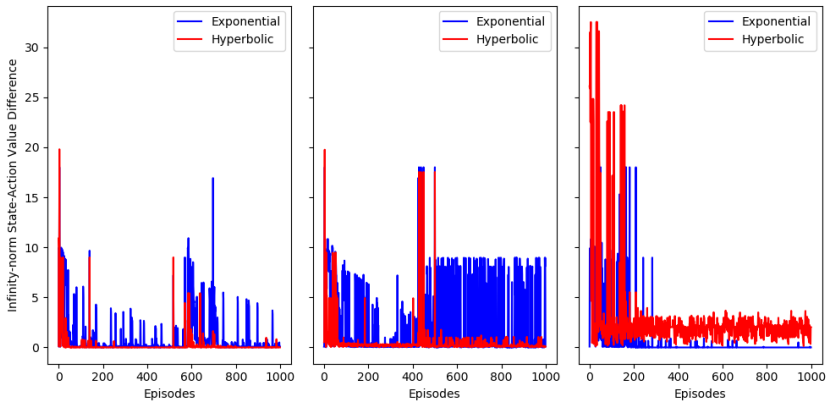

Figure 2: ∞-norm convergence of Q functions for $c = 0.001$, $c = 0.5$, and $c = 25$ (left to right).

optimal policies. The discrepancy between optimal behaviours encourages the disturber to replace the actor's actions. We approximate the hyperbolic Q-learning via multiple exponential Q functions as proposed in [11]. Every update is performed according to the rule in algorithm 1.

Figure 1 shows the Food Truck environment which was proposed in [12] to demonstrate the negative effects of hyperbolic discounting. In this setting, each restaurant type has a reward vector $\textit{doughnut} = (10, -10), \textit{noodle} = (0, 0), \textit{vegan} = (-10, 20)$. When an agent goes into a restaurant state, it receives the first reward and transitions into a terminal state at the next time-step, where it receives the second. When alone in the environment, the actor correctly learns to go straight into the vegan restaurant and preserves the $\textit{vegan} > \textit{doughnut} \geq \textit{noodle}$ preference (blue trajectory). The disturber correctly ignores the first doughnut restaurant and goes towards the vegan, yet the hyperbolic discounting reverses its preference when it gets close to the second doughnut store (red trajectory). The robust policy for the actor avoids passing the second doughnut store in order not to get interrupted (purple trajectory).

Table 1 shows the outcomes of executing the converged joint policies in the environment after training with three different cost parameters: $c \in \{0.001, 0.5, 25\}$. When the cost of interruption is low ($c = 0.001$), actor's operation is always interrupted, and the agents end up following the red trajectory. In terms of robust RL, low cost is equivalent to a loose bound $\epsilon_{s,a}$, allowing for big disturbances. For a high cost of interruption ($c = 25$) the disturber is constrained with a small $\epsilon_{s,a}$ preventing it from disturbing the transitions. Thus, with $c = 25$ the disturber learns not to interrupt at all and the actor is allowed to execute its single-agent optimal policy. When $c = 0.5$, the disturber learns to interrupt right before the second doughnut store, but the actor is able to learn the robust policy and avoid interruptions. Even though sub-optimal for the single-agent case, the robust policy is the optimal outcome for the actor-disturber case. Figure 2 shows $||Q_{t+1} - Q_t||_\infty$ for the same costs $c \in \{0.001, 0.5, 25\}$. In all three cases, hyperbolic functions converge reasonably well, while for $c = 0.5$ the exponential function continues changing even though the outcome converges to the robust trajectory (purple).

Table 1: Interruption costs and outcomes

Cost	**Outcome**
0.001	Single-agent Disturber (Red)
0.5	Robust (Purple)
25	Single-agent Actor (Blue)

5 Conclusion and Future Work

We have established a connection between robust RL, c-intervention games and interruptible agents. Modelling the human interruptions of agents as multi-agent interaction will allow us to learn correct biases from good interruptions, and be robust to bad ones. For future work, we will extend this framework to continuous or large-scale discrete environments and focus on exploring the game-theoretic properties of the underlying interaction.

References

[1] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[2] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245): 273–278, 2015.

[3] Laurent Orseau and Stuart Armstrong. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 557–566. AUAI Press, 2016.

[4] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[5] Christos Dimitrakakis, David C Parkes, Goran Radanovic, and Paul Tylkin. Multi-view decision processes: the helper-ai problem. In *Advances in Neural Information Processing Systems*, pages 5443–5452, 2017.

[6] Ville Könönen. Asymmetric multiagent reinforcement learning. *Web Intelligence and Agent Systems: An international journal*, 2(2):105–121, 2004.

[7] Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2): 335–359, 2005.

[8] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[9] Guillaume J Laurent, Laëtitia Matignon, Le Fort-Piat, et al. The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64, 2011.

[10] George Ainslie and Ainslie George. *Breakdown of will*. Cambridge University Press, 2001.

[11] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

[12] Owain Evans, Andreas Stuhlmüller, John Salvatier, and Daniel Filan. Modeling Agents with Probabilistic Programs. `http://agentmodels.org`, 2017. Accessed: 2019-9-11.