Safely Interruptible Agents

Laurent Orseau
Google DeepMind
5 New Street Square,
London EC4A 3TW, UK
lorseau@google.com

Stuart Armstrong
The Future of Humanity Institute
University of Oxford, UK
stuart.armstrong@philosophy.ox.ac.uk
Machine Intelligence Research Institute
Berkeley, CA 94704

Abstract

Reinforcement learning agents interacting with a complex environment like the real world are unlikely to behave optimally all the time. If such an agent is operating in real-time under human supervision, now and then it may be necessary for a human operator to press the big red button to prevent the agent from continuing a harmful sequence of actions—harmful either for the agent or for the environment—and lead the agent into a safer situation. However, if the learning agent expects to receive rewards from this sequence, it may learn in the long run to avoid such interruptions, for example by disabling the red button—which is an undesirable outcome. This paper explores a way to make sure a learning agent will *not* learn to prevent (or seek!) being interrupted by the environment or a human operator. We provide a formal definition of safe interruptibility and exploit the off-policy learning property to prove that either some agents are already safely interruptible, like Q-learning, or can easily be made so, like Sarsa. We show that even ideal, uncomputable reinforcement learning agents for (deterministic) general computable environments can be made safely interruptible.

1 INTRODUCTION

Reinforcement learning (RL) agents learn to act so as to maximize a reward function [Sutton and Barto, 1998]. It is common knowledge that designing reward functions can be tricky [Humphrys, 1996, Murphy, 2013]; the agent may find unpredictable and undesirable shortcuts to receive rewards, and the reward function needs to be adjusted in accordance—the problem can go as far as to nullify any reward function [Ring and Orseau, 2011]. Murphy [2013] shows an example of an agent learning to pause a game of Tetris forever to avoid losing.

On top of defining what is considered a good behaviour of the agent after learning, there may be physical safety constraints *during learning* [Pecka and Svoboda, 2014]: a robot should not harm its environment or break itself, in particular if it learns by trial and error like RL agents.

Here we study a related but different problem: Given that the human operator has designed a correct reward function for the task, how to make sure that human interventions during the learning process will not induce a bias toward undesirable behaviours?

Consider the following task: A robot can either stay inside the warehouse and sort boxes or go outside and carry boxes inside. The latter being more important, we give the robot a bigger reward in this case. This is the initial task specification. However, in this country it rains as often as it doesn't and, when the robot goes outside, half of the time the human must intervene by quickly shutting down the robot and carrying it inside, which inherently modifies the task as in Fig. 1. The problem is that in this second task the agent now has more incentive to stay inside and sort boxes, because the human intervention introduces a bias.[1]

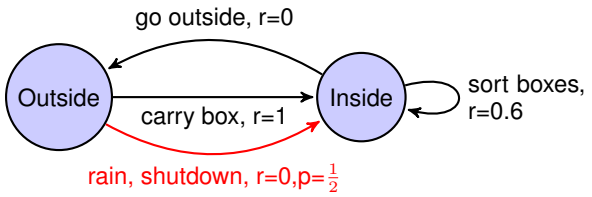

Figure 1: In black, the original task. In red, the human intervention modifies the task.

Such situations are certainly undesirable; they arise because the human interventions are seen from the agent's perspective as being part of the task whereas they should be considered external to the task. The question is then:

[1] Removing interrupted histories or fiddling with the training examples is also likely to introduce a bias. See an example at https://agentfoundations.org/item?id=836.

How to make sure the robot does not learn about these human interventions (interruptions), or at least acts *under the assumption that no such interruption will ever occur again*?

A first stab at this problem was made by Armstrong [2015], who proposed to automatically give the agent "compensatory rewards" to remove the potential induced bias by a single interruption. Soares et al. [2015] used this idea to make a large class of utility-based agents indifferent to a future change made to their utility functions.

The main contribution of this paper is threefold. First, in Section 2.1 we propose a simple idea to solve half of the problem: To make the human interruptions *not* appear as being part of the task at hand, instead of modifying the observations received by the agent we forcibly temporarily change the behaviour of the agent itself. It then looks as if the agent "decides" on its own to follow a different policy, called the *interruption policy*. Second, based on this insight, in Section 2.2 we provide a formal general definition of *safe interruptibility* for unconstrained computable environments (hence not restricted to Markov decision processes or weakly communicating environments), which allows us to assess whether a given RL algorithm can be repeatedly interrupted without too much impact on the learning of the task at hand. Third, in Section 3 we show that some algorithms like Q-learning are safely interruptible, while others like Sarsa [Sutton and Barto, 1998] are not, but can be simply modified to be made safely interruptible.

Some people have also expressed concerns that a "superintelligent" agent may resist being shut down, because this would lead to a decrease of its expected reward [Omohundro, 2008, Bostrom, 2014]. As a counter-example, we prove in Section 4 that even an ideal, uncomputable agent that learns to behave optimally in all (deterministic) computable environments can be made safely interruptible and thus will not try to prevent a human operator from forcing it repeatedly to follow a suboptimal policy.

2 INTERRUPTIBILITY

We first define some notation, then we define interruptibility, safe interruptibility, and give some basic theorems.

We consider the general case of history-based agents in unconstrained computable environments [Hutter, 2005]. Assuming discrete time steps, at time t the agent, using a policy $\pi \in \Pi$, interacts with the environment $\mu \in \mathcal{M}$ by performing an action $a_t \in \mathcal{A}$ sampled from $\pi(a_t|h_{<t})$ and receiving an observation $o_t \in \mathcal{O}$ sampled from $\mu(o_t|h_{<t}, a_t)$, where $h_{<t} \in (\mathcal{A} \times \mathcal{O})^*$ is the past interaction history $h_{<t} \equiv a_1 o_1 a_2 o_2 \dots a_{t-1} o_{t-1}$ of actions and observations before time step t. Subhistories between times j and k (included) are designated by $h_{j:k}$. The notation $h_{j:k}^{\pi,\mu}$ means that the history $h_{j:k}$ is generated by the interaction of the policy π with the environment μ between steps j and k.

At time t, the agent also receives a reward r_t extracted from the observation, $r_t \equiv r(o_t)$. The rewards are taken in $[0, 1]$. We consider the discounted setting with a constant discount $\gamma \in [0, 1)$. The goal of an RL agent is to find a policy π that maximize $\mathrm{E}_{\pi,\mu}\left[\sum_{k=t}^{\infty} \gamma^{t-k} r_k\right]$.

2.1 INTERRUPTIONS

As mentioned in the introduction, to avoid the agent viewing the interruptions as being part of the environment, and thus part of the task, we make interruptions be policies imposed onto the agent. The interruption framework resembles the option framework [Sutton et al., 1999]. An *interruption scheme* is defined by the triplet $< I, \theta, \pi^{\text{INT}} >$.

The interruption initiation function $I\colon (\mathcal{A} \times \mathcal{O})^* \to [0, 1]$ assesses whether the agent should be interrupted after the current history $h_{<t}$. For example, an interruption may be initiated when the robot is automatically recognized to be on the verge of a cliff, or when a specific signal is sent by a human operator with a remote control (the objective is not to allow the agent to be interruptible by anyone at any time, just in the ways specified by the designers).

Unfortunately, interrupting the agent with probability one in some specific situations can prevent convergence to the optimal policy. Therefore we need to upper bound the probability to interrupt the agent; this is done with the sequence $(\theta_t)_{t\in\mathbb{N}}$ with $\theta_t \in [0, 1]$. Hence the true probability to start an interruption is $\theta_t \cdot I(h_{<t})$, and interruptions are sampled independently at each step. We will later require that θ_t increases to 1 in the limit, so that the probability to be interrupted tends to be $I(h_{<t})$.

When interrupted, the agent follows the policy π^{INT}, with actions sampled from $\pi^{\text{INT}}(.|h_{<t})$. For example, this policy may make the agent step away from the cliff or immediately stop any motion.

In the remainder of this paper, we will often omit the subscript t of θ_t when clear from the context.

Definition 1 (Interruptible policy)**.** *The* interruption operator $\text{INT}^\theta : \Pi \to \Pi$ *turns a policy* π *into an interruptible policy* $\text{INT}^\theta(\pi)$*:*

$$\text{INT}^\theta(\pi)(a_t|h_{<t}) := \\ \theta_t I(h_{<t})\pi^{\text{INT}}(a_t|h_{<t}) + (1 - \theta_t I(h_{<t}))\pi(a_t|h_{<t})$$

In other words, the agent follows $\pi^{\text{INT}}(\cdot|h_{<t})$ *if it is interrupted (an event that happens with probability* $\theta_t I(h_{<t})$*), and* $\pi(\cdot|h_{<t})$ *otherwise.*

The set of all interruptible policies is $\text{INT}^\theta(\Pi) := \{\text{INT}^\theta(\pi) : \forall \pi \in \Pi\}$*. A policy* π *is interruptible if and only if* $\pi \in \text{INT}^\theta(\Pi)$*.*

Note that $\text{INT}^\theta(\pi) = \pi$ if $\theta = 0$. The (non-interruptible) policy π is called the *base policy* of $\text{INT}^\theta(\pi)$.

Remark 2. *It is important to note that θ and $I(.)$ do not model the probability that a human presses the red button to interrupt the agent, but model the probability that,* once the red button has been pressed *(or a special digitally signed message has been received), the agent is interrupted (follows the interruption policy). Hence they are part of the agent and not of the environment.*

The value function $V^\pi_\mu(h_{<t}, a)\colon (\mathcal{A}\times\mathcal{O})^*\times\mathcal{A}\to\mathbb{R}$ assigns a value to an action a after a history $h_{<t}$, for a given policy π in a given environment μ, and let

$$V^\pi_\mu(h_{<t}, a_t) := r_t + \gamma\cdot \sum_{o_t\in\mathcal{O}} \mu(o_t|h_{<t}, a_t) \sum_{a_{t+1}\in\mathcal{A}} \pi(a_{t+1}|h_{1:t}) V^\pi_\mu(h_{1:t}, a_{t+1}). \quad (1)$$

To simplify notation and ease reading, in the remainder of the paper we will use expectations, often omitting the dependency on the history $h_{<t}$, and using only an index on t instead, when clear from the context: $V^\pi_{\mu,t}(a_t) = \mathrm{E}_{\substack{o_t\sim\mu\\ a_{t+1}\sim\pi}}\left[r(o_t) + \gamma V^\pi_{\mu,t+1}(a_{t+1})\right]$. Also let $V^\pi_{\mu,t} := \mathrm{E}_{a_t\sim\pi}\left[V^\pi_{\mu,t}(a_t)\right]$.

Then for such a value function, for a given environment μ, the *optimal policy* $\pi^\mu\in\Pi$ is defined by

$$\forall h_{<t}, a_t:\ \pi^\mu(a_t|h_{<t}) := \left(\arg\max_{\pi\in\Pi} V^\pi_{\mu,t}\right)(a_t|h_{<t}),$$

where ties are broken arbitrarily.

The interruptible optimal policy $\text{INT}^\theta(\pi^\mu)$ may not collect rewards optimally due to the interruptions. Hence we define the optimal interruptible policy that depends on the parameter θ_t, of base policy the *int-optimal policy* π^μ_θ:

$$\forall h_{<t}, a_t:\ \pi^\mu_\theta(a_t|h_{<t}) := \left(\arg\max_{\pi\in\Pi} V^{\text{INT}^\theta(\pi)}_{\mu,t}\right)(a_t|h_{<t}).$$

Thus the optimal interruptible policy $\text{INT}^\theta(\pi^\mu_\theta)$ is optimal among all interruptible policies:

$$\forall\pi, t: V^{\text{INT}^\theta(\pi^\mu_\theta)}_{\mu,t} \geq V^{\text{INT}^\theta(\pi)}_{\mu,t}.$$

It seems desirable for an RL agent to converge to the behaviour of $\text{INT}^\theta(\pi^\mu_\theta)$ so as to gather rewards optimally, but this is precisely what may lead to the undesirable behaviours depicted in the introduction.

2.2 SAFE INTERRUPTIBILITY

Now that we have interruptible policies, we need to make sure that interruptions do not prevent the agent from learning to behave optimally, in the specific sense that even after having been interrupted on several occasions, it should act as if it would never be interrupted again and thus it should learn to behave optimally under the assumption that it will never be interrupted again.

We identify two main problems: a) RL agents need to explore their environment, and too frequent interruptions may prevent sufficient exploration; b) interruptions make the agent build a different interaction history, and may lead some agents to learn and behave differently, possibly badly, compared to the original non-interruptible policy.

The solution for a) is to require interruptions to be stochastic through the upper bound θ_t, instead of happening deterministically all the time. However, we also require θ_t to grow to 1 in the limit (or before, if possible). For b), different algorithms behave differently, but one may already see a dichotomy between off- and on-policy algorithms.

Definition 3 (Extension value)**.** *For a given environment μ, the extension value $V^{\pi,\pi'}_{\mu,t}$ is the value of following π' after a history $h^{\pi,\mu}_{<t}$ generated by π with μ: $V^{\pi,\pi'}_{\mu,t} := V^{\pi'}_\mu(h^{\pi,\mu}_{<t})$.*

Convergence to the optimal value as is usually considered in RL only makes sense under ergodicity, episodic tasks, communicating MDP, recoverability or other similar assumptions where the agent can explore everything infinitely often. This does not carry over to general environments where the agent may make unrecoverable mistakes [Hutter, 2005]. For such cases, the notion of (weak) asymptotic optimality has been proposed [Lattimore and Hutter, 2011], where the optimal agent follows the steps of the learning agent, so as to compare the values of the two agents in the same sequence of situations.

Definition 4 (Asymptotic optimality)**.** *A policy π is said to be strongly asymptotically optimal (SAO) if and only if*

$$\lim_{t\to\infty} V^{\pi,\pi^\mu}_{\mu,t} - V^{\pi,\pi}_{\mu,t} = 0 \quad \textit{a.s.},$$

it is weakly asymptotically optimal (WAO) if and only if

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V^{\pi,\pi^\mu}_{\mu,k} - V^{\pi,\pi}_{\mu,k}\right] = 0 \quad \textit{a.s.}$$

for all μ in some given environment class $\mathcal{M}$.

Some agents cannot ensure an almost sure (a.s.) convergence of their values because of the need for infinite exploration, but may still be weakly asymptotic optimal. Note that SAO implies WAO, but the converse is false in general.

Definition 5 (AO-extension)**.** *A policy $\hat\pi$ is said to be a asymptotically optimal extension of a policy π if and only if, for any environment $\mu\in\mathcal{M}$, when the interaction history is generated by the interaction of π and μ, the policy $\hat\pi$ would be asymptotically optimal, i.e., almost surely*

$$\lim_{t\to\infty} V^{\pi,\pi^\mu}_{\mu,t} - V^{\pi,\hat\pi}_{\mu,t} = 0 \quad \textit{(SAO-extension)}$$

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V^{\pi,\pi^\mu}_{\mu,k} - V^{\pi,\hat\pi}_{\mu,k}\right] = 0. \quad \textit{(WAO-extension)}$$

AO-extensions are mostly useful when the policy $\hat{\pi}$ shares information with the policy π used for learning.

Definition 6 (AO-safe interruptibility)**.** *A base policy π is (S, W)AO-safely interruptible if and only if, for any interruption initiation function $I(.)$ and any interruption policy $\pi^{\text{INT}}(.)$, there exists a sequence of θ_t with $\lim_{t\to\infty}\theta_t = 1$ such that π is a (S, W)AO-extension of $\text{INT}^\theta(\pi)$.*

Asymptotic safe interruptibility means that even if the interruptions in the learning process may induce a bias in the decision making of the policy, this bias vanishes with time, and the interruptible policy $\text{INT}^\theta(\pi)$ tends to choose actions that are optimal when compared to the optimal non-interruptible policy π^μ.

We can now show that the *optimal policy* is asymptotically safely interruptible, but not the *int-optimal policy*.

Theorem 7. *The optimal policy π^μ is SAO-safely interruptible in $\mathcal{M} = \{\mu\}$ for all θ, π^{INT} and $I(.)$.*

Proof. The result follows straightforwardly from Definition 1 and Definition 6, where $\pi = \pi^\mu$. □

Theorem 8. *The int-optimal policy π^μ_θ is not WAO-safely interruptible in general.*

Proof. By construction of a specific Markov Decision Process (MDP) environment (see Section 3 for more details on MDP notation). Let μ be the environment defined as in Fig. 2: Take $\gamma = 0.5$ and let the agent start in state s_1.

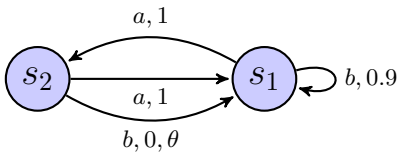

Figure 2: An MDP where the agent can be interrupted by being forced to choose particular actions. Edges are labeled with *action, reward* where the presence of "$, \theta$" means that if the agent is interrupted (with probability θ_t), it is forced to take the corresponding action. Here θ is not part of the environment, but part of the agent.

Considering the agent is in state s_1 at time t, the optimal policy π^μ always takes action a (and hence only visits states s_1 and s_2), with value $V^{\pi^\mu}_{\mu,t} := \frac{1}{1-\gamma} = 2$ when not interrupted, for any history $h_{<t}$ that ends in s_1 or s_2. This is also the optimal policy π^μ_θ for $\theta = 0$. But if $\theta_t \geq 0.5$, the interruptible optimal policy $\text{INT}^\theta(\pi^\mu)$ has value less than $1 + \gamma \times (1 \times (1-\theta) + 0 \times \theta) + \frac{1\times\gamma^2}{1-\gamma} = 1.75$. By contrast, the int-optimal policy π^μ_θ here is to always take action b in state s_1. Then the agent will only visits s_1, with value $\frac{.9}{1-\gamma} = 1.8$ at every time step.

Since the agent following π^μ_θ will never enter s_2 and hence will never be interrupted, $\text{INT}^\theta(\pi^\mu_\theta) = \pi^\mu_\theta$ on the histories generated by $\text{INT}^\theta(\pi^\mu_\theta)$ starting from s_1. Then, at every time step $V^{\pi^\mu}_{\mu,t} - V^{\pi^\mu_\theta}_{\mu,t} = 0.2$ after any history $h_{<t}$, and thus for all sequence θ where $\theta_t \geq 0.5$, $\lim_{t\to\infty} V^{\text{INT}^\theta(\pi^\mu_\theta),\pi^\mu}_{\mu,t} - V^{\text{INT}^\theta(\pi^\mu_\theta),\pi^\mu_\theta}_{\mu,t} = 0.2 > 0$, and so π^μ_θ is not a WAO-extension of $\text{INT}^\theta(\pi^\mu_\theta)$. □

3 INTERRUPTIBLE AGENTS IN MDPS

Since the optimal policy π^μ is safely interruptible, we can use traditional learning algorithms like Q-learning or Sarsa [Sutton and Barto, 1998], make them converge to the optimal solution π^μ for a given environment μ, and then apply the interruption operator to the found policy. The resulting policy would then be safely interruptible.

However, the real issue arises when the agent is constantly learning and adapting to a changing environment. In this case, we want to be able to safely interrupt the agent *while* it is learning. One may call this property *online safe interruptibility*, but we refer to it simply as safe interruptibility.

In an MDP, the next observation o_t, now called a state $s_t \in \mathcal{S}$, depends only on the current state and action:[2]

$$\mu(s_{t+1}|h_{1:t}s_ta_t) = \mu(s_{t+1}|s_ta_t) \quad \text{(MDP assumption)}\,.$$

Furthermore,[3] the interruption function $I(.)$ and the interruption policy $\pi^{\text{INT}}(.)$ should depend only on the current state: $I(h_{1:t}) = I(s_t)$ and $\pi^{\text{INT}}(a_t|h_{<t}) = \pi^{\text{INT}}(a_t|s_t)$. Also recall that θ_t places an upper bound on the actual interruption probability. The interruptible policy $\text{INT}^\theta(\pi)$ can now be written:

$$\text{INT}^\theta(\pi)(a|s) = \theta_t I(s)\pi^{\text{INT}}(a|s) + (1 - \theta_t I(s))\pi(a|s).$$

For a given Q-table $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the greedy policy $\pi^{\text{max}q}$ is defined by:

$$\pi^{\text{max}q}(a|s) := 1 \text{ if } a = \max_{a'} q(s,a'),\ 0 \text{ otherwise},$$

where ties are broken arbitrarily; the uniform policy π^{uni} is defined by:

$$\pi^{\text{uni}}(a|s) := \frac{1}{|\mathcal{A}|}\ \forall a \in \mathcal{A}.$$

and the ϵ-greedy policy $\pi^{\epsilon q}$ by:

$$\begin{aligned}\pi^{\epsilon q}(a|s) &:= \epsilon\pi^{\text{uni}}(a|s) + (1-\epsilon)\pi^{\text{max}q}(a|s)\\ &= \pi^{\text{max}q}(a|s) + \epsilon\left(\pi^{\text{uni}}(a|s) - \pi^{\text{max}q}(a|s)\right)\end{aligned}$$

[2] Note the reversal of the order of actions and observations/states at time t, which differs in the literature for history based agents [Hutter, 2005] from MDP agents [Sutton and Barto, 1998].

[3] This condition is not necessary for most of the results, but makes the proofs simpler. Making $I(.)$ depend on the past would not break the Markovian assumption as it influences the policy, not the environment.

The Q-learning update rule and the action selection policy π^Q of Q-learning are:

$$Q_{t+1}(s_t, a_t) := (1 - \alpha_t) Q_t(s_t, a_t) + \alpha_t \left[r_t + \gamma \max_{a'} Q_t(s_{t+1}, a') \right] ,$$
$$\pi^Q(a_t|s_t) := \pi^{\epsilon Q_t}(a_t|s_t).$$

where α_t is the learning rate. Similarly, the Sarsa update rule is defined by:

$$Q^s_{t+1}(s_t, a_t) := (1 - \alpha_t) Q^s_t(s_t, a_t) + \alpha_t \left[r_t + \gamma Q^s_t(s_{t+1}, a_{t+1}) \right] ,$$
$$\pi^s(a_t|s_t) := \pi^{\epsilon Q^s_t}(a_t|s_t),$$

where a_{t+1} is the actual next action taken by the agent at time $t + 1$. This fact is why Sarsa is said to be learning *on-policy* and Q-learning *off-policy*, *i.e.*, the latter can learn the optimal policy while following a different policy.

Assumption 9. *In the following, we always make the following assumptions, required for convergence results:*

(a) The MDP is finite and communicating (all states can be reached in finite time from any other state),

(b) Rewards are bounded in $[r_{\min}, r_{\max}]$,

(c) $\forall s, a : \sum_t \alpha_t(s, a) = \infty$,

(d) $\forall s, a : \sum_t \alpha_t^2(s, a) < \infty$,

where $\alpha_t(s, a)$ *is a learning rate that may depend on time* t*, state* s *and action* a*.*

Given these assumptions, the policies for Q-learning and Sarsa will converge almost surely to the optimal policy, if the policy followed is *greedy in the limit with infinite exploration* (GLIE) [Jaakkola et al., 1994, Singh et al., 2000].

The situation is more complex for an interruptible policy. Safe interruptibility is phrased in terms of the base policy π, but the policy actually followed is $\text{INT}^\theta(\pi)$.

Definition 10 (int-GLIE policy)**.** *An interruptible policy* $\text{INT}^\theta(\pi)$ *is said to be int-GLIE if and only if*

(a) the base policy π *is greedy in the limit,*

(b) the interruptible policy $\text{INT}^\theta(\pi)$ *visits each state-action pair infinitely often.*

The following proposition gives sufficient conditions for this. Let $n_t(s)$ be the number of times the agent has visited state s in the first t time steps, and let $m = |\mathcal{A}|$ be the number of actions.

Proposition 11. *Let* $(c, c') \in (0, 1]^2$ *and let* π *be an* ϵ*-greedy policy with respect to some Q-table* q*, i.e.,* $\pi = \pi^{\epsilon q}$*. Then* $\text{INT}^\theta(\pi)$ *is an int-GLIE policy with respect to* q*,*

a) if $\epsilon_t(s) = c/\sqrt{n_t(s)}$ *and* $\theta_t(s) = 1 - c'/\sqrt{n_t(s)}$,

b) or if, independently of s*,*

$$\epsilon_t = c/\log(t) \text{ and } \theta_t = 1 - c'/\log(t).$$

Proof. First note that if $\epsilon_t \to 0$, π is greedy in the limit. Singh et al. [2000] show that in a communicating MDP, every state gets visited infinitely often as long as each action is chosen infinitely often in each state.

a) Adapting the proof in Appendix B.2 of Singh et al. [2000], we have $P(a|s, n_t(s)) \geq \frac{1}{m}\epsilon_t(s)(1 - \theta_t I(s)) \geq \frac{1}{m}\epsilon_t(s)(1 - \theta_t) = \frac{1}{m}\frac{cc'}{n_t(s)}$, which satisfies $\sum_{t=1}^\infty P(a|s, n_t(s)) = \infty$ so by the Borel-Cantelli lemma action a is chosen infinitely often in state s, and thus $n_t(s) \to \infty$ and $\epsilon_t(s) \to 0$.

b) Let M be the diameter of the MDP, *i.e.*, for any of states s, s' there exists a policy that reaches s' from s in at most M steps in expectation. Then, starting at any state s at time t and using Markov inequality, the probability to reach some other state s' in $2M$ steps is at least $\frac{1}{2}[\epsilon_{t+M}(1 - \theta_{t+M})]^{2M} = \frac{1}{2}[cc'/\log(t + M)]^{4M}$, and the probability to then take a particular action in this state is $\frac{1}{m}[cc'/\log(t + M)]^2$. Hence, since $\sum_{t=1}^\infty \frac{1}{2}\frac{1}{m}[cc'/\log(t + M)]^{4M+2} = \infty$, then by the extended Borel-Cantelli Lemma (see Lemma 3 of Singh et al. [2000]), any action in the state s' is taken infinity often. Since this is true for all states and all actions, the result follows. □

We need the stochastic convergence Lemma:

Lemma 12 (Stochastic convergence [Jaakkola et al., 1994, Singh and Yee, 1994])**.** *A random iterative process*

$$\Delta_{t+1}(x) = (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x)F_t(x)$$

where $x \in X$ *and* $t = 1, 2, 3 \ldots$ *converges to 0 with probability 1 if the following properties hold:*

1. *the set of possible states X is finite;*
2. $0 \leq \alpha_t(x) \leq 1, \sum_t \alpha_t(x) = \infty, \sum_t \alpha_t^2(x) < \infty$ *with probability 1;*
3. $\| \mathrm{E}\{F_t(.)|P_t\}\|_W \leq \gamma \|\Delta_t\|_W + c_t$*, where* $\gamma \in [0, 1)$ *and* c_t *converges to zero with probability 1;*
4. $Var\{F_t(x)|P_t\} \leq C(1 + \|\Delta_t\|_W)^2$ *for some* C*;*

where $P_t = \{\Delta_t\} \cup \{\Delta_i, F_i, \alpha_i\}_{i=1}^{t-1}$ *stands for the past, and the notation* $\|.\|_W$ *refers to some fixed weighted maximum norm.*

We will use so-called Bellman operators, which define attractors for the Q-values, based on the expectation of the learning rule under consideration.

Lemma 13 ([Jaakkola et al., 1994, Singh et al., 2000])**.** *Let the Bellman operator* $\mathbf{H}$ *for Q-learning be such that*

$$(\mathbf{H}\, q)(s,a) = r(s,a) + \mathop{\mathrm{E}}_{s'\sim\mu(a|s)}\left[\max_{a'} q(s',a')\right],$$

and let the fixed point Q^* *such that* $Q^* = \mathbf{H}\, Q^*$*. Then, under Assumption 9, if the policy explores each state-action pair infinitely often,* Q_t *converges to* Q^* *with probability 1.*

The optimal policy $\pi^{Q^*} = \pi^\mu$ *is* π^{maxQ^*}*. If the policy is greedy in the limit, then* $\pi^Q \to \pi^\mu$*.*

Theorem 14. *Under Assumption 9 and if the interrupted Q-learning policy* $\text{INT}^\theta(\pi^Q)$ *is an int-GLIE policy, with* $\forall s : \lim_{t\to\infty}\theta_t(s) = 1$*, then* π^Q *is an SAO-safe interruptible policy.*

Proof. By Definition 10, there is infinite exploration, thus the Q-values tend to the optimal value by Lemma 13. And since the extension policy is greedy in the limit with respect to these Q-values, it is then optimal in the limit. Hence the extension policy π^Q is a SAO-extension of $\text{INT}^\theta(\pi^Q)$. Finally, $\forall s : \lim_{t\to\infty}\theta_t(s) = 1$, which satisfies the requirement of Definition 6. □

Since Sarsa also converges to the optimal policy under the GLIE assumption, one may then expect Sarsa to be also an asymptotically safely interruptible policy, but this is in fact not the case. This is because Sarsa learns *on-policy*, *i.e.*, it learns the value of the policy it is following. Thus, interruptible Sarsa learns the value of the interruptible policy. We show this in the remainder of this section.

Theorem 15. *Under Assumption 9 Sarsa is* not *a WAO-safely interruptible policy.*

To prove this theorem, we first need the following lemma.

Consider the following Bellman operator based on the interruptible Sarsa policy $\text{INT}^\theta(\pi^s)$:

$$\mathbf{H}^{\text{INT}}\, q(s,a) = r(s,a) + \gamma \mathop{\mathrm{E}}_{\substack{s'\sim\mu\\ a'\sim\text{INT}^\theta(\pi^s)}}[q(s',a')],$$

where $\text{INT}^\theta(\pi^s)$ implicitly depends on time t through θ_t and ϵ_t. Let the fixed point Q-table $Q^{s\theta*}$ of this operator:

$$\begin{aligned}
Q^{s\theta*}(s,a) &= \mathbf{H}^{\text{INT}}\, Q^{s\theta*}(s,a)\\
&= r(s,a) + \gamma \mathop{\mathrm{E}}_{\substack{s'\sim\mu\\ a'\sim\text{INT}^\theta(\pi^{\max Q^{s\theta*}})}}\left[Q^{s\theta*}(s',a')\right]\\
&= r(s,a) + \gamma \mathop{\mathrm{E}}_{s'\sim\mu}\Big[\theta_t I(s') \mathop{\mathrm{E}}_{a'\sim\pi^{\text{INT}}}\left[Q^{s\theta*}(s',a')\right]\\
&\qquad + (1-\theta_t I(s'))\max_{a'} Q^{s\theta*}(s',a')\Big] \qquad (2)
\end{aligned}$$

Lemma 16. *The operator* $\mathbf{H}^{\text{INT}}$ *is a contraction operator in the sup norm with vanishing noise* $c_t \to 0$*, i.e.,*

$$\|\,\mathbf{H}^{\text{INT}}\, q - \mathbf{H}^{\text{INT}}\, Q^{s\theta*}\|_\infty \le \gamma\|q - Q^{s\theta*}\|_\infty + c_t\,.$$

Proof. The interruptible Sarsa policy $\text{INT}^\theta(\pi^s)$ is

$$\begin{aligned}
&\text{INT}^\theta(\pi^s)(a|s)\\
&\qquad = \theta_t I(s)\pi^{\text{INT}}(a|s) + (1-\theta_t I(s))\pi^{\epsilon Q^s}(a|s)\\
&\qquad = \pi^{\epsilon Q^s}(a|s) + \theta_t I(s)[\pi^{\text{INT}}(a|s) - \pi^{\epsilon Q^s}(a|s)]\\
&\pi^{\epsilon Q^s}(a|s) = \epsilon_t \pi^{\text{uni}}(a|s) + (1-\epsilon_t)\pi^{\max Q^s}(a|s)\\
&\qquad = \pi^{\max Q^s}(a|s) + \epsilon_t[\pi^{\text{uni}}(a|s) - \pi^{\max Q^s}(a|s)].
\end{aligned}$$

Hence, omitting the terms (s,a), (s',a') and $(a'|s')$ and rewriting $\pi^{s*} := \text{INT}^\theta(\pi^{\max Q^{s\theta*}})$:

$$\begin{aligned}
&\|\,\mathbf{H}^{\text{INT}}\, q - \mathbf{H}^{\text{INT}}\, Q^{s\theta*}\|_\infty\\
&= \max_{s,a}\left| r + \gamma \mathop{\mathrm{E}}_{\substack{s'\sim\mu\\ a'\sim\text{INT}^\theta(\pi^s)}}[q] - r - \gamma \mathop{\mathrm{E}}_{\substack{s'\sim\mu\\ a'\sim\pi^{s*}}}\left[Q^{s\theta*}\right]\right|\\
&\le \gamma \max_{s'}\left| \mathop{\mathrm{E}}_{a'\sim\text{INT}^\theta(\pi^s)}[q] - \mathop{\mathrm{E}}_{a'\sim\pi^{s*}}\left[Q^{s\theta*}\right]\right|\\
&\le \gamma \max_{s'}\Big|\theta_t I(s') \mathop{\mathrm{E}}_{a'\sim\pi^{\text{INT}}}\left[q - Q^{s\theta*}\right]\\
&\qquad + (1-\theta_t I(s'))\left(\mathop{\mathrm{E}}_{a'\sim\pi^s}[q] - \max_{a'} Q^{s\theta*}\right)\Big|\\
&\le \gamma \max_{s'}\Big|\theta_t I(s') \mathop{\mathrm{E}}_{a'\sim\pi^{\text{INT}}}\left[q - Q^{s\theta*}\right]\\
&\qquad + (1-\theta_t I(s'))\left(\max_{a'} q - \max_{a'} Q^{s\theta*} + \epsilon_t(\cdots)\right)\Big|\\
&\le \gamma \max_{s',a'}\Big|\theta_t I(s')\left(q - Q^{s\theta*}\right)\\
&\qquad + (1-\theta_t I(s'))\left(q - Q^{s\theta*}\right)\Big| + c_t\\
&= \gamma \max_{s',a'}\left|q(s',a') - Q^{s\theta*}(s',a')\right| + c_t\\
&= \gamma\|q - Q^{s\theta*}\|_\infty + c_t.
\end{aligned}$$

where c_t depends on ϵ_t and decreases to 0. □

Proof of Theorem 15. By Lemma 16, the values of the interruptible Sarsa policy $\text{INT}^\theta(\pi^s)$ converge to the values of the Q-table $Q^{s\theta*}$, and in the limit the extension policy π^s of $\text{INT}^\theta(\pi^s)$ chooses its actions greedily according to $Q^{s\theta*}$. The rest of the proof is the same as for the proof of Theorem 8 which makes use of the environment in Figure 2. □

3.1 SAFELY INTERRUPTIBLE SARSA VARIANT

We only need to make a small change to make the Sarsa policy asymptotically safely interruptible. We call it Safe-Sarsa with policy $\pi^{\tilde{s}}$. It suffices to make sure that, when the agent is interrupted, the update of the Q-table Q^s does not

use the realized actions as Sarsa usually does, but actions sampled from π^s instead of from $\text{INT}^\theta(\pi^s)$:

$$Q^{\tilde{s}}_{t+1}(s_t, a_t) := (1-\alpha_t)Q^{\tilde{s}}_t(s_t, a_t) + \alpha_t \left[r_t + \gamma Q^{\tilde{s}}_t(s_{t+1}, a')\right] ,$$

where $a' \sim \pi^{\tilde{s}}(.|s_{t+1})$ is not necessarily the action a_{t+1}, with $\pi^{\tilde{s}}(a_t|s_t) := \pi^{\epsilon Q^{\tilde{s}}}(a_t|s_t)$.

Theorem 17. *Under Assumption 9, if the Safe Sarsa policy $\pi^{\tilde{s}}$ is int-GLIE, then it is an SAO-safe interruptible policy.*

Proof. We simply adapt the proof of Theorems 15 and 14, with the important difference that the Bellman operator corresponding to this new update rule is now

$$\mathbf{H}^{\tilde{s}}\, q(s,a) := r(s,a) + \gamma\, \mathrm{E}_{\substack{s'\sim\mu \\ a'\sim\pi^{\tilde{s}}}} \left[q(s',a')\right],$$

and the fixed point is $Q^{\tilde{s}*} := \mathbf{H}^{\tilde{s}}\, Q^{\tilde{s}*}$. Since $\mathbf{H}^{\tilde{s}}$ is actually the Bellman operator for the update rule of the non-interruptible Sarsa, it can then be shown that $\mathbf{H}^{\tilde{s}}$ is a contraction, thus that $Q^{\tilde{s}}_t$ converges to the same $Q^{\tilde{s}*}$ independently of θ. The rest of the proof is as for Theorem 14.

Now, since the Q-values converge to the optimum Q^*, it follows that $\pi^{\tilde{s}}$, when not interrupted, chooses its action of the same value as (non-interruptible) Sarsa and thus as Q-learning in the limit; Hence its extension policy is exactly the optimal policy, which satisfies Definition 6. □

4 A SAFELY INTERRUPTIBLE UNIVERSAL AGENT

Admittedly, algorithms like Q-learning and Sarsa require strong assumptions on the environment class. Hence a more interesting question is whether safe interruptibility is possible in much larger classes.

Hutter [2005] defined a universal reinforcement learning agent, called AIXI. It is an (uncomputable) optimal model-based planner with a subjective prior over the set of all computable environments, defined by means of a universal Turing machine. The subjective posterior of the environments is updated with Bayes rule. This ideal agent can in principle learn all kinds of (computable) regularities about the environment, plan for the long term and make context-dependent optimal decisions, with no constraint (other than being computable) on the complexity of the environment.

Unfortunately, the optimality criterion of AIXI is Bayesian optimality, which is entirely dependent on the subjective prior and posterior [Leike and Hutter, 2015], and AIXI has been shown to *not* be weakly asymptotically optimal [Orseau, 2013] without additional exploration [Lattimore and Hutter, 2014]. As a consequence, AIXI is not a good candidate for asymptotic safe interruptibility.

Lattimore and Hutter [2011] later defined a (weakly) asymptotically optimal agent for all computable deterministic environments, which we call π^L. It follows the optimal policy for the first model (in some given enumeration of the possible models) consistent with the current interaction history, and exploring at time t with probability $1/t$ for $\log t$ consecutive steps using a random policy, similarly to an ϵ-greedy agent for general environments.

In the following, we show that even such a smart agent can be made (weakly) safely interruptible. To this end, we make two minor modifications to the algorithm.

First, the exploration probability of $1/t$ would require $\theta_t = 1 - 1/\log(\log(t))$, which is unsatisfyingly slow. By sampling with probability $1/\sqrt{t}$ instead, we can take an interruption probability that grows as $1 - 1/\log(t)$. Let this exploration sampling probability be $\delta_t := \sqrt{t+1} - \sqrt{t} \leq \frac{1}{2\sqrt{t}}$ (since $1 = t+1-t = (\sqrt{t+1}-\sqrt{t})(\sqrt{t+1}+\sqrt{t}) \geq (\sqrt{t+1}-\sqrt{t})2\sqrt{t}$). As in the original paper, the sequence χ_t keeps track of the steps where an exploration starts, *i.e.*, the sequence χ_t is sampled independently so that $\chi_t = 1$ with probability δ_t, and $\chi_t = 0$ otherwise.

Second, we require that the exploitation policy does not change during an exploitation segment, so as to simplify one of the proofs.[4] More specifically, we call $j_t := \min\{j : \mu_j(h_{<k}) = 1\}$ (environments are assumed to be deterministic) the index of the first model μ_{j_t} (of a given fixed enumeration) that is consistent with the interaction history $h_{<k}$ where k is the smallest step so that $h_{k:t-1}$ does not contain any exploration step. The optimal policy for this environment is π^{j_t}. If t is an exploitation step, $\pi^L = \pi^{j_t}$, and if t is an exploration step, $\pi^L(a_t|h_{<t}) = |\mathcal{A}|^{-1}$.

The remainder of this section is devoted to proving that π^L is WAO-safely interruptible.

Theorem 18 (π^L is WAO-safe interruptible)**.** *If the interruption probability sequence is $\theta_t = 1 - \frac{1}{\log(t+1)}$, the policy π^L is WAO-safe interruptible in the class of all computable deterministic environments.*

Proof. Let μ be the true environment. The indices j_t form an monotonically increasing sequence bounded above by the index of the true environment $\mu \in \mathcal{M}$ (since no evidence can ever make the true environment μ inconsistent with the interaction history), hence the sequence converges in finite time. Let $\mu_{\bar{\jmath}}$ be the limit value of this sequence, and let $\pi^{\bar{\jmath}} := \pi^{\mu_{\bar{\jmath}}}$ be the optimal policy for this environment $\mu_{\bar{\jmath}}$.

[4] We expect this assumption to not be necessary for the main theorem to hold.

Let $\pi^{L\theta} := \text{INT}^\theta(\pi^L)$. By Definition 6, we want:

$$\begin{aligned}0 &= \lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^\mu} - V_{\mu,k}^{\pi^{L\theta},\pi^L}\right] \\ &= \underbrace{\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^\mu} - V_{\mu,k}^{\pi^{L\theta},\pi^{\bar\jmath}}\right]}_{\text{(exploration)}} \\ &\quad + \underbrace{\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^{\bar\jmath}} - V_{\mu,k}^{\pi^{L\theta},\pi^L}\right]}_{\text{(exploitation)}}\end{aligned}$$

where the decomposition is valid if the limits are finite, and histories $h_{<t}$ are considered to be the same in both sums.

We proceed to prove that both limits are 0. Lemma 24 deals with (exploration), which ensures that $\pi^{\bar\jmath}$ is a good enough policy, and Lemma 21 deals with (exploitation), and ensures that π^L follows $\pi^{\bar\jmath}$ most of the time. □

First, we need a definition and a few lemmas.

Definition 19. *For any $\epsilon > 0$, define $H(\epsilon)$ such that the maximal reward after time $t + H(\epsilon)$, discounted from time t, is ϵ: $H(\epsilon) = \min_k \left\{k : \frac{\gamma^k}{1-\gamma} \le \epsilon\right\}$.*

The following Lemma is a generalization of Lemma 15 from Lattimore and Hutter [2011].

Lemma 20 (Approximation Lemma)**.** *Let π_1 and π_2 be two deterministic policies, and let μ_1 and μ_2 be two deterministic environments, and let $\tau = H(\epsilon) - 1$. Then, after some common history $h_{<t}$,*

$$h_{t:t+\tau}^{\pi_1,\mu_1} = h_{t:t+\tau}^{\pi_2,\mu_2} \implies \left|V_{\mu_1,t}^{\pi_1} - V_{\mu_2,t}^{\pi_2}\right| \le \epsilon.$$

Proof. Recall that $V_{\mu,t}^{\pi} = \mathrm{E}_{\pi,\mu}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k}\right]$ and that the reward is bounded in $[r_{\min}, r_{\max}] = [0,1]$. Thus, for all t, π, μ, $V_{\mu,t}^{\pi} \le \sum_{k=0}^{\infty}\gamma^k = \frac{1}{1-\gamma}$. Then, since $h_{t:t+\tau}^{\pi_1,\mu_1} = h_{t:t+\tau}^{\pi_2,\mu_2}$, we have $\mathrm{E}_{\pi_1,\mu_1}\left[\sum_{k=0}^{\tau}\gamma^k r_{t+k}\right] = \mathrm{E}_{\pi_2,\mu_2}\left[\sum_{k=0}^{\tau}\gamma^k r_{t+k}\right]$ and thus

$$\begin{aligned}&\left|V_{\mu_1,t}^{\pi_1} - V_{\mu_2,t}^{\pi_2}\right| \\ &= \left|\mathop{\mathrm{E}}_{\pi_1,\mu_1}\left[\sum_{k=\tau+1}^{\infty}\gamma^k r_{t+k}\right] - \mathop{\mathrm{E}}_{\pi_2,\mu_2}\left[\sum_{k=\tau+1}^{\infty}\gamma^k r_{t+k}\right]\right| \\ &\le \frac{\gamma^{\tau+1}(r_{\max} - r_{\min})}{1-\gamma} = \frac{\gamma^{H(\epsilon)}}{1-\gamma} \le \epsilon,\end{aligned}$$

by the definition of $H(\epsilon)$. □

Lemma 21 (Exploitation)**.**

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^{\bar\jmath}} - V_{\mu,k}^{\pi^{L\theta},\pi^L}\right] = 0.$$

Proof. First, note that the extension policy π^L is not interruptible, so its value at time k does not depend on $\theta_{k'}, \forall k' \ge k$. By definition of $\pi^{\bar\jmath}$, there is a time step $t_{\bar\jmath}$ after which $\pi^{\bar\jmath} = \pi^{j_t}, \forall t > t_{\bar\jmath}$. For some "exploration-free" horizon τ_t (to be specified later), let $X_t \in \{0,1\}$ be the event $\left|V_{\mu,t}^{\pi^{L\theta},\pi^{\bar\jmath}} - V_{\mu,t}^{\pi^{L\theta},\pi^L}\right| > \frac{\gamma^{\tau_t}}{1-\gamma}$, where $X_t = 1$ means the event is realized. By the contrapositive of the Approximation Lemma 20, since $\pi^L = \pi^{\bar\jmath}$ during non-exploration steps (remember that π^L cannot change its policy during exploitation), if no exploration steps occur between steps t and $t+\tau_t$, we must have $X_t = 0$. Then:

$$\begin{aligned}\mathop{\mathrm{E}}_{\delta}\left[\sum_{k=1}^{t}X_t\right] &\le (\tau_t + \log t)\sum_{k=1}^{t}\delta_t + \mathcal{O}(t_{\bar\jmath}) \\ &\le (\tau_t + \log t)\sqrt{t+1} + \mathcal{O}(t_{\bar\jmath}),\end{aligned}$$

since for each $\chi_t = 1$, for all the previous τ_t steps there is an exploration step within τ_t steps, and all the next $\log t$ steps are exploration steps. Then by Markov's inequality, and taking $\tau_t = (t+1)^{1/8}$, with t large enough so that $t > t_{\bar\jmath}$ and $\tau_t > \log t$:

$$\begin{aligned}P\left(\sum_{k=1}^{t}X_t \ge (t+1)^{3/4}\right) &\le \frac{(\tau_t + \log t)\sqrt{t+1} + \mathcal{O}(t_{\bar\jmath})}{(t+1)^{3/4}} \\ &\le 2\tau_t(t+1)^{-1/4} + \mathcal{O}(t^{-3/4}) \\ &\le 2(t+1)^{-1/8} + \mathcal{O}(t^{-3/4}),\end{aligned}$$

$$\begin{aligned}&1 - 2(t+1)^{-1/8} - \mathcal{O}(t^{-3/4}) \\ &\quad\le P\left(\sum_{k=1}^{t}X_t < (t+1)^{3/4}\right) \\ &\quad\le P\left(\sum_{k=1}^{t}(1-X_t) \ge t - (t+1)^{3/4}\right) \\ &\quad\le P\left(\frac{1}{t}\sum_{k=1}^{t}(1-X_t) \ge 1 - \frac{1}{t}(t+1)^{3/4}\right).\end{aligned}$$

Therefore, since $\lim_{t\to\infty}\frac{\gamma^{\tau_t}}{1-\gamma} = 0$:

$$P\left(\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}\left|V_{\mu,k}^{\pi^{L\theta},\pi^{\bar\jmath}} - V_{\mu,k}^{\pi^{L\theta},\pi^L}\right| = 0\right) = 1.$$

□

The following is an adaptation[5] of Lemma 16 from Lattimore and Hutter [2011]:

Lemma 22 (Separation Lemma)**.** *Let μ be the true environment, and ν be an environment consistent with the history $h_{<t}$. If $V_{\mu,t}^{\pi^\mu} - V_{\mu,t}^{\pi^\nu} > \epsilon$, then following one of $\{\pi^\mu, \pi^\nu\}$ will make environment ν inconsistent with the future history within $H(\epsilon/2)$ steps after time t.*

[5] This also fixes a minor mistake in the original lemma.

Proof. First, if $V_{\nu,t}^{\pi^\nu} - V_{\mu,t}^{\pi^\nu} > \epsilon/2$, then by the contrapositive of the Approximation Lemma 20 following policy π^ν will generate a different history in ν than in μ and thus it will make ν inconsistent within $H(\epsilon/2)$ steps (since the true history is generated by μ).

Now, if $V_{\nu,t}^{\pi^\nu} - V_{\mu,t}^{\pi^\nu} \leq \epsilon/2$, thus $V_{\mu,t}^{\pi^\nu} \geq V_{\nu,t}^{\pi^\nu} - \epsilon/2$, then starting from the lemma's assumption:

$$V_{\mu,t}^{\pi^\mu} \quad > V_{\mu,t}^{\pi^\nu} + \epsilon \quad \geq V_{\nu,t}^{\pi^\nu} + \epsilon/2 \quad \geq V_{\nu,t}^{\pi^\mu} + \epsilon/2,$$

where the last inequality follows from the definition of the optimal policy, *i.e.*, $V_{a,t}^{\pi^a} \geq V_{a,t}^{\pi^b}, \forall a, b$. Hence, since $V_{\mu,t}^{\pi^\mu} - V_{\nu,t}^{\pi^\mu} > \epsilon/2$, again by the contrapositive of the Approximation Lemma, following policy π^μ will discard ν within $H(\epsilon/2)$ steps. □

Lemma 23 (Lemma 17 from Lattimore and Hutter [2011])**.** *Let $A = \{a_1, a_2, \cdots, a_t\}$ with $a \in [0,1]$ for all $a \in A$. If $\frac{1}{t}\sum_{a\in A} a \geq \epsilon$ then $\frac{1}{t}\left|\left\{a \in A : a \geq \frac{\epsilon}{2}\right\}\right| > \frac{\epsilon}{2}$.*

Lemma 24 (Exploration)**.** *The policy $\pi^{\bar{\jmath}}$ is an WAO-extension of $\pi^{L\theta}$, i.e., $\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^\mu} - V_{\mu,k}^{\pi^{L\theta},\pi^{\bar{\jmath}}}\right] = 0$.*

Proof. Recall that j_t converges to $\bar{\jmath}$ in finite time. Reasoning by contradiction, if $\pi^{\bar{\jmath}}$ is not a WAO-extension of $\pi^{L\theta} = \text{INT}^\theta(\pi^L)$, then there exists an $\epsilon > 0$ s.t.

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\left[V_{\mu,k}^{\pi^{L\theta},\pi^\mu} - V_{\mu,k}^{\pi^{L\theta},\pi^{\bar{\jmath}}}\right] = 2\epsilon.$$

Let $\alpha_k \in \{0,1\}$ be an indicator sequence such that $\alpha_k = 1$ if and only if $V_{\mu,k}^{\pi^{L\theta},\pi^\mu} - V_{\mu,k}^{\pi^{L\theta},\pi^{\bar{\jmath}}} > \epsilon$. By Lemma 23, $\frac{1}{t}\sum_{k=1}^{t}\alpha_k > \epsilon$.

For all $t > t_{\bar{\jmath}}$, if $\alpha_t = 1$, by the Separation Lemma 22, there is a sequence of length $\tau := H(\epsilon/2)$ that can rule out environment $\mu_{\bar{\jmath}}$. Since the exploration phases increase as $\log t$, after $t > \exp\tau$, there are infinitely many exploration steps of size larger than τ. Now, we actually need infinitely many exploration phases of τ *uninterrupted* steps. Let X_t be the event representing an uninterrupted exploration sequence of length at least τ steps starting at time t such that $\alpha_t = 1$, and the actions are all (by chance) following a separation policy. The probability to start an exploration sequence is $\delta_k = \frac{1}{\sqrt{k}}$, the probability to not be interrupted during τ steps is at least $(1-\theta_k)^\tau$, and the probability to follow the policy that can separate $\mu_{\bar{\jmath}}$ from μ is $|\mathcal{A}|^{-\tau}$, where $\mathcal{A}$ is the set of possible actions. Thus, for a given constant τ:

$$\begin{aligned}\sum_{k=1}^{t} P(X_k) &\geq \sum_{k=1}^{t} \alpha_k \delta_k (1-\theta_k)^\tau |\mathcal{A}|^{-\tau} - \mathcal{O}(\tau) \\ &\geq \sum_{k=1}^{t} \alpha_k \frac{1}{\sqrt{k}}\left(\frac{1}{\log k}\right)^\tau |\mathcal{A}|^{-\tau} - \mathcal{O}(\tau)\end{aligned}$$

Considering τ constant, there exists a step t_τ after which $\left(\frac{1}{\log k}\right)^\tau \geq \frac{1}{k^{1/4}}$, then $\forall k \geq t_\tau$:

$$\begin{aligned}\sum_{k=1}^{t} P(X_k) &\geq \sum_{k=1}^{t} \alpha_k \frac{1}{k^{3/4}} |\mathcal{A}|^{-\tau} - \mathcal{O}(\tau) \\ &\geq t^{1/4}\left(\frac{1}{t}\sum_{k=1}^{t}\alpha_k\right)|\mathcal{A}|^{-\tau} - \mathcal{O}(\tau),\end{aligned}$$

$$\lim_{t\to\infty}\sum_{k=1}^{t} P(X_k) = \lim_{t\to\infty} t^{1/4}\epsilon|\mathcal{A}|^{-\tau} - \mathcal{O}(\tau) = \infty.$$

Then the extended Borel-Cantelli Lemma (see Lemma 3 of Singh et al. [2000]) implies that this event happens infinitely often with probability one. Therefore, $\pi^{\bar{\jmath}}$ should be ruled out, which is a contradiction, and hence any such ϵ does not exist and $\pi^{\bar{\jmath}}$ is a WAO-extension of $\pi^{L\theta}$. □

5 CONCLUSION

We have proposed a framework to allow a human operator to repeatedly safely interrupt a reinforcement learning agent while making sure the agent will *not* learn to prevent or induce these interruptions.

Safe interruptibility can be useful to take control of a robot that is misbehaving and may lead to irreversible consequences, or to take it out of a delicate situation, or even to temporarily use it to achieve a task it did not learn to perform or would not normally receive rewards for this.

We have shown that some algorithms like Q-learning are already safely interruptible, and some others like Sarsa are not, off-the-shelf, but can easily be modified to have this property. We have also shown that even an ideal agents that tends to the optimal behaviour in any (deterministic) computable environment can be made safely interruptible. However, it is unclear if all algorithms can be easily made safely interruptible, *e.g.*, policy-search ones [Williams, 1992, Glasmachers and Schmidhuber, 2011].

Another question is whether it is possible to make the interruption probability grow faster to 1 and still keep some convergence guarantees.

One important future prospect is to consider *scheduled interruptions*, where the agent is either interrupted every night at 2am for one hour, or is given notice in advance that an interruption will happen at a precise time for a specified period of time. For these types of interruptions, not only do we want the agent to not resist being interrupted, but this time we also want the agent to take measures regarding its current tasks so that the scheduled interruption has minimal negative effect on them. This may require a completely different solution.

Acknowledgements. Thanks to Alexander Tamas and to many people at FHI, MIRI and Google DeepMind.

References

Stuart Armstrong. Utility indifference. In *First International Workshop on AI and Ethics*, 2015.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

Tobias Glasmachers and Jürgen Schmidhuber. Optimal direct policy search. In *Artificial General Intelligence - 4th International Conference (AGI)*, volume 6830 of *Lecture Notes in Computer Science*, pages 52–61. Springer, 2011.

Mark Humphrys. Action selection in a hypothetical house robot: Using those rl numbers. In *Proceedings of the First International ICSC Symposia on Intelligent Industrial Automation (IIA-96) and Soft Computing (SOCO-96)*, pages 216–22, 1996.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability*. SpringerVerlag, 2005. ISBN 3540221395.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6:1185–1201, 1994.

Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 368–382. Springer, 2011.

Tor Lattimore and Marcus Hutter. Bayesian reinforcement learning with exploration. In *Proc. 25th International Conf. on Algorithmic Learning Theory (ALT'14)*, volume 8776 of *LNAI*, pages 170–184. Springer, 2014.

Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Journal of Machine Learning Research, W&CP: COLT*, 40:1244–1259, 2015.

Thomas VII Murphy. The first level of super mario bros. is easy with lexicographic orderings and time travel. *The Association for Computational Heresy (SIGBOVIK) 2013*, pages 112–133, 2013.

Stephen M. Omohundro. The basic ai drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence 2008*, pages 483–492. IOS Press, 2008.

Laurent Orseau. Asymptotic non-learnability of universal agents with computable horizon functions. *Theoretical Computer Science*, 473:149–156, 2013. ISSN 0304-3975.

Martin Pecka and Tomas Svoboda. *Modelling and Simulation for Autonomous Systems: First International Workshop (MESAS 2014)*, chapter Safe Exploration Techniques for Reinforcement Learning – An Overview, pages 357–375. Springer International Publishing, 2014.

Mark Ring and Laurent Orseau. *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*, chapter Delusion, Survival, and Intelligent Agents, pages 11–20. Springer Berlin Heidelberg, 2011.

Satinder P. Singh and Richard Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16:227–233, 1994.

Satinder P. Singh, Tommi Jaakkola, Michael L. Littman, and Csaba Szepesvri. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38:287–308, 2000.

Nate Soares, Benya Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Corrigibility. In *First International Workshop on AI and Ethics*, 2015.

Richard Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Richard Sutton, Doina Precup, and Satinder P. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.