Reinforcement Learning in Newcomblike Environments

James Bell* jbell@posteo.net
Linda Linsefors* linda.linsefors@gmail.com
Caspar Oesterheld* caspar.oesterheld@duke.edu
Joar Skalse* joar.skalse@cs.ox.ac.uk

December 13, 2020

Abstract

Newcomblike decision problems have been studied extensively in the Decision Theory literature, but they have so far been largely absent in the Reinforcement Learning literature. In this paper we study value-based reinforcement learning algorithms in the Newcomblike setting, and answer some of the fundamental theoretical questions about the behaviour of such algorithms in these environments. We show that a value-based reinforcement learning agent cannot converge to a policy that is not *ratifiable*. This gives us a powerful tool for reasoning about the limit behaviour of agents – for example, it lets us show that there are Newcomblike environments in which a reinforcement learning agent cannot converge to any optimal policy. We show that a ratifiable policy always exists in our setting, but that there are cases in which a reinforcement learning agent normally cannot converge to it (and hence cannot converge at all). We also prove several results about the possible limit behaviours of agents in cases where they do not converge to any policy.

1 Introduction

In reinforcement learning, the Markov Decision Process (MDP) is the canonical formalisation of a decision problem. However, there are broad classes of decision problems that cannot be satisfactorily formalised as MDPs. Consider the following two examples:

Newcomb's Problem (Nozick, 1969): There are two boxes in front of you; one opaque box, and one transparent box, and you can see that the transparent box contains $1,000. You can choose to either take only the opaque box, or to take both the opaque box and the transparent box. The boxes have been placed in this room by an agent who can predict your policy; if he believes that you will take only the opaque box then he has put $1,000,000 in the opaque box, and if he believes that you will take both boxes then he has left the opaque box

*All authors contributed equally.

empty. Do you take one box, or two?

Death in Damascus (Gibbard and Harper, 1976): Death will come for you tomorrow. You can choose to stay in Damascus (where you are currently) or you can flee to Aleppo. If you are in the same city as Death tomorrow, you will die. Death has already decided which city he will go to — however, he can predict your policy, and has decided to go to the city where he believes that you will be tomorrow. Do you stay in Damascus, or flee to Aleppo?

These two problems are examples of *Newcomblike decision problems*. A decision problem is Newcomblike if the actions of the decision maker can provide evidence about things that they cannot causally influence (or, more precisely, if they provide *more* evidence than what is provided by their causal impact). For example, in Newcomb's Problem the action taken by the decision maker provides evidence about the content of the opaque box, even though it cannot causally influence the content of the opaque box. If a situation is Newcomblike then this is typically (but not necessarily) because the procedure that the decision maker uses to make decisions is available to other actors. It is often not possible to formalise Newcomblike problems as MDPs.

Newcomblike problems are well-studied in the field of Decision Theory, but they have so far received little direct attention in the Reinforcement Learning literature. However, Newcomblike problems may be relatively common for reinforcement learning agents deployed in certain kinds of environments. For example, an agent whose source code is available to other actors can easily find itself in Newcomblike situations (cf. the literature on program equilibriums, e.g. Tennenholtz, 2004; Oesterheld, 2019b). Newcomblike decision problems can also occur when agents can encounter copies of themselves. For example, an agent who is playing the Prisoner's Dilemma against an identical copy of itself is in a situation that is very closely analogous to Newcomb's Problem (Brams, 1975; Lewis, 1979). More generally, game-theoretic situations can in some cases be Newcomblike (cf. Gauthier, 1989, Section XI). We believe that if an agent is deployed in a complex, open-ended environment that the designer of the agent does not fully control then this environment may in many cases be Newcomblike (Cavalcanti, 2010, Section 5; Oesterheld, 2019a, Section 1; Conitzer, 2019). It is therefore important to have a good theoretical understanding of how different algorithms will behave in such cases.

In this paper we analyse how value-based model-free reinforcement learning agents behave when they are placed in Newcomblike environments. In Section 2 we demonstrate that such algorithms can only converge to a policy that is *ratifiable* – that is, to a policy π for which all actions taken by π have optimal expected reward when following π.

In Section 3, we discuss the convergence properties of agents in Newcomblike situations, and show that there are cases where value-based agents must fail to converge.

When policies do not converge, then sometimes the action frequencies converge nonetheless. In Section 4, we establish some conditions on any action frequency that an agent could converge to. We show that there are decision problems and agents where even the action frequencies do not converge.

1.1 Newcomblike Decision Processes

MDPs are unable to model many Newcomblike decision problems in a satisfactory way. We therefore propose an alternative framework: let a *Newcomblike Decision Process (NDP)* be a tuple $\langle S, A, T, R, \gamma \rangle$ where

- S is a finite set of *states*;
- A is a finite set of *actions*;
- $T : S \times A \times (S \rightsquigarrow A) \rightsquigarrow S$ is a nondeterministic *transition function*;
- $R : S \times A \times S \times (S \rightsquigarrow A) \rightsquigarrow \mathbb{R}$ is a nondeterministic *reward function*, which we assume to be bounded; and
- $\gamma \in [0, 1)$ is a *discount factor*.

A policy $\pi : S \rightsquigarrow A$ is a function that nondeterministically maps states to actions. We use $\pi(a \mid s)$ to denote the probability of taking action a in state s while following the policy π. T and R are functions from states, actions, and policies. In other words, they allow the outcome of a decision to depend on the distributions from which the agent draws its actions, rather than just the state and the action that is in fact taken. Also note that $T(s, a, \pi)$ and $R(s, a, s', \pi)$ are defined even if $\pi(a \mid s) = 0$. We say that an NDP is a *Bandit NDP* if it has only one state. We will sometimes use $R(s, a, \pi)$ as a shorthand for $R(s, a, T(s, a, \pi), \pi)$, and we will sometimes omit the state from T, R, and π for Bandit NDPs. Moreover, we normally let $\gamma = 0$ for Bandit NDPs.

This framework makes it easier to formalise Newcomblike problems. For example, Newcomb's Problem can be formalised as the following Bandit NDP:

- $S = \{s\}$
- $A = \{a_{\text{one-boxing}}, a_{\text{two-boxing}}\}$
- $R(a_{\text{one-boxing}}, \pi) = \begin{cases} 0 & \text{with probability} \quad \pi(a_{\text{two-boxing}}) \\ 10 & \text{w.p.} \quad \pi(a_{\text{one-boxing}}) \end{cases}$
  $R(a_{\text{two-boxing}}, \pi) = \begin{cases} 5 & \text{w.p.} \quad \pi(a_{\text{two-boxing}}) \\ 15 & \text{w.p.} \quad \pi(a_{\text{one-boxing}}) \end{cases}$

We here suppose that the content of the transparent box is worth a utility of 5, and that content of the opaque box is worth either a utility of 10 or 0. We also assume that the predictor chooses whether or not to put any money in the opaque box by sampling from the policy of the decision maker, instead of predicting its actions perfectly.[1]

We say that an NDP is *continuous* if T and R are continuous in the policy. In this paper we work mainly with continuous NDPs. This is in part because it is technically convenient, and in part because we believe it is likely to be the case in many situations.[2]

[1] In most versions of Newcomb's Problem, the predictor directly predicts the action of the decision maker. However, this version of the problem can be modeled as a regular MDP – after all, the transition probabilities of an MDP could represent non-causal dependencies. The key difference between NDPs and MDPs is therefore not Newcomblikeness *per se*, but rather the dependence of the transition probability on the policy (be that dependence causal or not). However, in practice, one would imagine that difference to be closely connected with issues of causality.

[2] For example, even if the environment has direct access to the source code of the agent,

1.2 Reinforcement Learning Agents

We consider value-based reinforcement learning agents. Such agents have two main components; a *Q-function* $S \times A \to \mathbb{R}$ that predicts the expected future discounted reward conditional on taking a particular action in a particular state, and a *bandit algorithm* that is used to select actions in each state based on the Q-function. Given a policy π, we use $q_\pi(a \mid s)$ to denote the true expected future discounted reward conditional on taking action a in state s while following the policy π (and conditional on all subsequent actions being chosen by π). A model-free agent will update Q over time to make it converge to q_π for some π. If Q is represented as a lookup table the agent is said to be *tabular*. The theoretical analysis in this paper will focus on tabular agents, but many of the results should apply to non-tabular agents.

The Q-values can be updated in different ways. One method is to use the update rule

$$\begin{aligned} Q_{t+1}(a_t \mid s_t) \leftarrow &(1 - \alpha_t(s_t, a_t))\, Q_t(a_t \mid s_t) + \\ &\alpha_t(s_t, a_t)(r_t + \gamma \max_a Q_t(a \mid s_{t+1})), \end{aligned}$$

where a_t is the action taken at time t, s_t is the state visited at time t, r_t is the reward obtained at time t, and $\alpha_t(s, a)$ is a learning rate. This update rule is known as *Q-learning* (Watkins, 1986). Another widely used update rule is

$$\begin{aligned} Q_{t+2}(a_t \mid s_t) \leftarrow &(1 - \alpha_t(s_t, a_t))\, Q_{t+1}(a_t \mid s_t) + \\ &\alpha_t(s_t, a_t)(r_t + \gamma Q_{t+1}(a_{t+1} \mid s_{t+1})). \end{aligned}$$

This update rule is known as *SARSA* (Rummery and Niranjan, 1994). Yet another common update rule is *Expected SARSA*, which replaces $Q_t(a_{t+1} \mid s_{t+1})$ with $\mathbb{E}_{a\sim\pi}[Q_t(a \mid s_{t+1})]$ in the SARSA update rule (van Seijen et al., 2009). For the purposes of this paper it will not matter significantly how the Q-values are computed, as long as it is the case that if an agent converges to a policy π in some NDP and explores infinitely often then Q converges to q_π. We will later see that this is the case for Q-learning, SARSA, and Expected SARSA.

There are also several different bandit algorithms. Two types of agents that are widely used in practice and that we will refer to throughout the paper are *softmax agents* and *ϵ-Greedy agents*. The policy of a softmax agent with a sequence of temperatures $\beta_t \in \mathbb{R}_+$ is given by:

$$\pi_t(a \mid s) = \frac{\exp(Q_t(a \mid s)/\beta_t)}{\sum_{a' \in A} \exp(Q_t(a' \mid s)/\beta_t)}.$$

Unless otherwise stated we assume that $\beta_t \to 0$. The policy of an ϵ-Greedy agent with a sequence of exploration probabilities $\epsilon_t \in [0, 1]$ is given by:

$$\pi_t(a \mid s) = \begin{cases} 1 - \epsilon_t & \text{if} \quad a = \arg\max_{a' \in A} Q_t(a' \mid s) \\ \epsilon_t / (|A| - 1) & \text{otherwise.} \end{cases}$$

Unless otherwise stated we assume that $\epsilon_t \to 0$. We assume that ϵ-Greedy breaks ties for argmax, so that there is always some $a \in A$ such that $\pi(a \mid s) = 1 - \epsilon_t$.

it may in general not be feasible to extract the precise action probabilities from the code. However, it is always possible to estimate the action probabilities by sampling. If this is done then T and R will depend continuously on the policy.

We say that an agent is *greedy in the limit* if with probability converging to 1 it takes an action that maximises Q, and we say that it *explores infinitely often* if it in the limit takes every action in every state infinitely many times. A softmax agent or ϵ-Greedy agent can be both greedy in the limit and explore infinitely often if β_t and ϵ_t go to 0 at the right rate.

1.3 Some Initial Observations

We here make some simple observations about the setting that we will use to prove and understand the results throughout this paper. First, note that a continuous NDP always has a policy π for which the expected discounted reward $\mathbb{E}[R \mid \pi]$ is maximised, since $\mathbb{E}[R \mid \pi]$ exists and is continuous in π, and since the set of possible policies is a compact set. Also note that an NDP in which T or R is discontinuous may not have any optimal policy.

We can also note that some NDPs (unlike MDPs) have no *deterministic* optimal policy. To see this, consider again the Death in Damascus problem. We can formalise this problem as an NDP $\langle S, A, T, R, \gamma \rangle$ in the following way: $S = \{s\}$, $A = \{a_{\text{Damascus}}, a_{\text{Aleppo}}\}$, and

$$R(a_{\text{Damascus}}, \pi) = \begin{cases} 0 & \text{w.p.} \quad \pi(a_{\text{Damascus}}) \\ 10 & \text{w.p.} \quad \pi(a_{\text{Aleppo}}) \end{cases}$$

$$R(a_{\text{Aleppo}}, \pi) = \begin{cases} 10 & \text{w.p.} \quad \pi(a_{\text{Damascus}}) \\ 0 & \text{w.p.} \quad \pi(a_{\text{Aleppo}}) \end{cases}$$

We here suppose that escaping Death is worth a utility of 10, and that meeting one's fate is worth a utility of 0. We also assume that Death chooses which city to go to by sampling from the same random policy as the agent (as opposed to predicting the agent's action perfectly, which is the case in the original version of the problem). In this NDP, the policy that goes to each city with equal probability outperforms all deterministic policies.

Note also that the Bellman optimality equation does not hold for NDPs. This equation states that for any MDP $\langle S, A, T, R, \gamma \rangle$ and any $s \in S$ we have that $v_{\pi^*}(s)$ equals

$$\max_{a \in A} \left(\sum_{s' \in S} P(T(s,a){=}s')(\mathbb{E}\left[R(s,a,s')\right] + \gamma v_{\pi^*}(s')) \right),$$

where π^* is the policy that maximises expected discounted reward, and $v_{\pi^*}(s)$ is the expected discounted future reward when visiting s and following π^* (Bellman, 1957). One might suppose that the same relationship should hold in NDPs, if T and R are parameterised by π^*. However, this is not the case. To see this, consider Newcomb's problem, as formalised in Section 1.1.

2 Ratifiability

If a reinforcement learning algorithm in the limit only takes the actions with the highest Q-values and it converges to some policy π_∞, then it is clear that all actions that π_∞ assigns positive probability to must have equal expected utility given π_∞. Otherwise, the Q-values would eventually reflect the differences in expected utility and the agent would move away from π_∞. Similarly,

if the algorithm explores sufficiently often, the actions that are taken with limit probability 0 cannot be better given π_∞ than those taken by π_∞. After all, if they were better, the agent would have eventually figured this out and assigned them large probability.

This condition on π_∞ resembles a well-known doctrine in philosophical decision theory: ratificationism (see Weirich, 2016, for an overview). One form of ratificationism is based on a distinction between a *decision* – what the agent chooses – and the *act* that is selected by that decision. Very roughly, ratificationism then states that a decision is rational only if the acts it selects have the highest expected utility given the decision. For instance, in Newcomb's Problem a decision to take one box would usually not be seen as ratifiable, because given that decision, the agent would rather perform the act of taking two boxes. In philosophical decision theory, concepts of causality are often invoked to formalise the difference between the decision, the act, and their respective consequences. Our setup, however, has such a differentiation built in: we will view the policy as the "decision" and the action sampled from it as the "act".

2.1 Strong Ratifiability

As hinted earlier, slightly different versions of the concept of ratifiability are relevant depending on how much exploration a learning algorithm guarantees. We start with the stronger version, which more closely resembles what philosophers mean when they speak about ratifiability.

Definition 1. *Let $M \subseteq S$ be a set of states. A policy π is strongly ratifiable on M if $\operatorname{supp}(\pi(\cdot \mid s)) \subseteq \arg\max_{a \in A} q_\pi(a \mid s)$ for all $s \in M$.*

In Newcomb's Problem there is only one strongly ratifiable policy, namely to take both boxes with probability 1. In Death in Damascus there is also just one strongly ratifiable policy, and that is to go to each city with probability 1/2. There can also be more than one strongly ratifiable policy. For example, if you play the Coordination Game of Table 1 against an opponent who samples his action from the same policy as you then there are three strongly ratifiable policies; to select action a with probability 1, to select action b with probability 1, and to select a with probability 1/3 and b with probability 2/3.

	a	b
a	2,2	0,0
b	0,0	1,1

Table 1: The Coordination Game

Theorem 2. *Let $\mathcal{A}$ be a model-free reinforcement learning agent, and let π_t and Q_t be $\mathcal{A}$'s policy and Q-function at time t. Let $\mathcal{A}$ satisfy the following in a given NDP $\langle S, A, T, R, \gamma \rangle$:*

- *$\mathcal{A}$ is greedy in the limit, i.e. for all $\delta > 0$, $\mathbb{P}(Q_t(\pi_t(s)) \leq \max_a Q_t(a \mid s) - \delta) \to 0$ as $t \to \infty$.*
- *$\mathcal{A}$'s Q-values are accurate in the limit, i.e. if $\pi_t \to \pi_\infty$ then $Q_t \to q_{\pi_\infty}$.*

Then if $\mathcal{A}$'s policy converges to π_∞ then π_∞ is strongly ratifiable on S.

In Appendix A we show that the Q-values of a tabular agent are accurate in the limit in any continuous NDP if the agent updates its Q-values with SARSA, Expected SARSA, or Q-learning, given that the agent explores infinitely often and uses appropriate learning rates. This means that such an agent can only converge to a (strongly) ratifiable policy, if it converges to any policy at all. Moreover, since we would expect most well-designed agents to have accurate Q-values in the limit, this result should apply very broadly. Using Kakutani's fixed-point theorem, it can be shown that every continuous NDP has a ratifiable policy.

Theorem 3. *Every continuous NDP has a strongly ratifiable policy.*

Of course, the fact that a ratifiable policy always exists does not necessarily mean that a reinforcement learning agent must converge to it — we will consider the question of whether or not this is the case in Section 3. It is also worth noting that a discontinuous NDP may not have any strongly ratifiable policy.

It is a topic of ongoing discussion among philosophical decision theorists whether (strong) ratifiability should be considered a normative principle of rationality, see Weirich (2016, Section 3.6) for details. In general, the policy π that maximises $\mathbb{E}[R \mid \pi]$ may or may not be ratifiable. For example, in Death in Damascus the optimal policy is to go to each city with equal probability, and this policy is also ratifiable, but in Newcomb's Problem the optimal policy is to take one box with probability 1, and this policy is not ratifiable. Hence there are NDPs in which most reinforcement learning agents cannot converge to the policy that maximises $\mathbb{E}[R \mid \pi]$. In fact, by scaling the payoffs in Newcomb's Problem we can obtain an NDP in which the only ratifiable policy is arbitrarily bad compared to the optimal policy.

2.2 Weak Ratifiability

We have seen that with sufficient exploration, the limit policy π_∞ (if it exists) must be strongly ratifiable. We will now show that even without infinite exploration, π_∞ must still satisfy a weaker notion of ratifiability.

Definition 4. *Let $M \subseteq S$ be a set of states. A policy π is weakly ratifiable on M if $q_\pi(a \mid s)$ is constant across $a \in \text{supp}(\pi(s))$ for all $s \in M$.*

Theorem 5. *Same conditions as theorem 2, but where $\mathcal{A}$'s Q-values are only required to be accurate in the limit for state-action pairs that $\mathcal{A}$ visits infinitely many times. Then π_∞ is weakly ratifiable on the set of states that are visited infinitely many times.*

What makes this a weak version of ratifiability is that it does not put any requirements on the expected utility of actions that π_∞ does not take, it merely says that all actions that π_∞ takes with positive probability must have the same (actual) q-value. As a special case, this means that all deterministic policies are weakly ratifiable. This result may therefore be of limited interest.

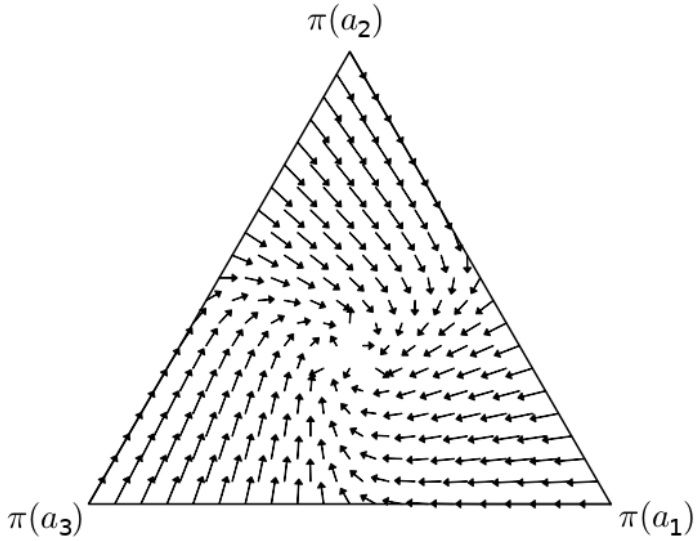

Figure 1: The triangle shows the space of possible policies in the Repellor Problem, parameterised by the probability they assign to each of the three actions. Plotted against this space is the expected direction in which a softmax agent would change its policy if playing a particular policy.

3 Non-Convergence of Policies

We have shown that most reinforcement learning algorithms can only converge to (strongly) ratifiable policies. We will now consider the question of whether or not they always converge to a policy at all. We find that this is not the case.

3.1 Theoretical Results

It should be clear that there are NDPs in which an ϵ-Greedy agent cannot converge to any policy. Most agents (including ϵ-Greedy agents) can only converge to strongly ratifiable policies. Moreover, in some NDPs all strongly ratifiable policies are mixed strategies that an ϵ-Greedy agent cannot express (Death in Damascus is an example of such an NDP). An ϵ-Greedy agent could therefore not possibly converge to any policy in these NDPs.

There are also NDPs in which a (slow-cooling) softmax agent cannot converge to any policy. As an example, consider a Bandit NDP with three actions a_1, a_2, a_3, and where the rewards $R(a_i, \pi)$ have expectations

$$\pi(a_{i+1}) + 4{\cdot}13^3{\cdot}\pi(a_i)\mathbb{1}\left[\forall j{:}\pi(a_j){\geq}\frac{1}{4}\right]\prod_j\left(\pi(a_j){-}\frac{1}{4}\right). \tag{1}$$

For $i = 3$, we here let $a_{i+1} = a_1$. We also require that the rewards are stochastic with a finite set of outcomes such that the empirical Q-values are never exactly equal between different actions. We call this the *Repellor Problem*.

Theorem 6. *Let $\mathcal{A}$ be an agent that plays the Repellor Problem, explores infinitely often, and updates its Q-values with a learning rate α_t that is constant*

across actions, and let π_t and Q_t be $\mathcal{A}$'s policy and Q-function at time t. Assume also that for $j \neq i$, if $\pi_t(a_i)$, $\pi_t(a_j)$ both converge to positive values, then

$$\frac{\pi_t(a_i) - \pi_t(a_j)}{Q_t(a_i) - Q_t(a_j)} \underset{a.s.}{\to} \infty \tag{2}$$

as $t \to \infty$. Then π_t almost surely does not converge as $t \to \infty$.

Line 2 is satisfied, for example, for softmax agents with a temperature converging to 0. Recall also that e.g. Q-learning and SARSA are equivalent for Bandit NDPs (if $\gamma = 0$).

3.2 Empirical Results

Empirically, it seems to be the case that softmax agents converge (to strongly ratifiable policies) in many NDPs, provided that the temperature decreases sufficiently slowly. To illustrate this we will use *Asymmetric Death in Damascus*, which is a version of the Death in Damascus problem where you assign a utility of 5 to dying in Aleppo. This version of the problem is due to Egan (2007), and is described by the decision matrix in Table 2. We use Asymmetric Death in Damascus, instead of the original version, to make it easier to distinguish between the case where an agent converges to the ratifiable policy and the case where it simply picks each action equally often as a default.

	Death goes to Damascus	Death goes to Aleppo
Stay in Damascus	0	10
Flee to Aleppo	10	5

Table 2: Asymmetric Death in Damascus

As before, we assume that Death chooses which city to go to by sampling from your policy. This NDP has only one (strongly) ratifiable policy, namely to go to Aleppo with probability 2/3 and Damascus with probability 1/3. Figure 2 shows the probability of converging to this policy with a softmax agent and a plot of the policy on one run. We can see that this agent reliably converges provided that the cooling is sufficiently slow.

However, there are also fairly natural games in which it seems like softmax agents cannot converge. The Repellor Problem was constructed to make Theorem 6 as easy as possible – a more natural example is *Loss-Averse Rock-Paper-Scissors* (LARPS), the problem of playing Rock-Paper-Scissors against an opponent that selects each action with the same probability as you, and where you assign utility 1 to a win, 0 to a draw, and -10 to a loss. This game is described by the payoff matrix in Table 3.

The only strongly ratifiable policy in LARPS is to take each action with equal probability, and hence we know that a softmax agent cannot converge to any policy other than this policy. However, this policy appears to be unstable. We thus conjecture that slow-cooling softmax agents do not converge in LARPS. We have unfortunately not been able to prove this formally, but Figure 3 presents some empirical data which corroborates the hypothesis.

	Rock	Paper	Scissors
Rock	0	-10	1
Paper	1	0	-10
Scissors	-10	1	0

Table 3: Loss-Averse Rock-Paper-Scissors

4 Convergence of Action Frequencies

We have seen that there are some NDPs in which some reinforcement learning algorithms cannot converge to any policy. But if they do not converge to any policy, what does their limit behaviour look like? This is the question that we will now consider. In particular, we will examine whether these algorithms must converge to taking each action with some limit frequency, and what sorts of frequencies they can converge to.

4.1 Possible Frequencies in the Bandit Case

In this section we establish a number of conditions that must be satisfied for a frequency to be a possible limit action frequency of a value-based agent. We consider agents that converge to deterministic behaviour (such as ϵ-Greedy agents). We limit our analysis to the Bandit case (with $\gamma = 0$).

Let $P_t^\Sigma \colon A \to [0,1]$ be the frequency with which each action in A is taken in the first t steps (for some agent and some Bandit NDP). Note that P_t^Σ is a random variable. By the law of large numbers,

$$P_t^\Sigma(a) - \frac{1}{t}\sum_{i=0}^{t} \pi_i(a) \underset{t\to\infty}{\to}_{\text{a.s.}} 0. \tag{3}$$

Let π_a be the policy that takes action a with probability 1, and let $q_a = q_{\pi_a}$.

Theorem 7. *Assume that there is some sequence of random variables $(\epsilon_t \geq 0)_t$ s.t. $\epsilon_t \underset{t\to\infty}{\to}_{a.s.} 0$ and for all $t \in \mathbb{N}$ it is*

$$\sum_{a^* \in \arg\max_a Q_t(a)} \pi_t(a^*) \geq 1 - \epsilon_t. \tag{4}$$

Let $P_t^\Sigma \to p^\Sigma$ with positive probability as $t \to \infty$. Then across all actions $a \in \text{supp}(p^\Sigma)$, $q_a(a)$ is constant.

This condition is vaguely analogous to weak ratifiability, and is proven in roughly the same way as Theorem 2.

Theorem 8. *Same assumptions as Theorem 7. If $|\text{supp}(p^\Sigma)| > 1$ then for all $a \in \text{supp}(p^\Sigma)$ there exists $a' \in A$ s.t. $q_a(a') \geq q_a(a)$.*

This condition is an instability condition. Say that multiple actions are taken with non-zero limit frequency, and that action a has the highest Q-value at time t. Then for other actions to be played with positive limit frequency, other actions must at some point be believed to be optimal again (since the probability of exploration goes to zero). Hence they cannot all be worse when explored while mainly playing a, since a could otherwise be played forever.

Theorem 9. *Same assumptions as Theorem 7. Let U be the Q-value $q_a(a)$ which (by Theorem 7) is constant across $a \in \text{supp}(p^\Sigma)$. For any $a' \in A - \text{supp}(p^\Sigma)$ that is played infinitely often, let frequency 1 of the exploratory plays of a' happen when playing a policy near elements of $\{\pi_a \mid a \in \text{supp}(p^\Sigma)\}$. Then there is $a \in \text{supp}(p^\Sigma)$ such that $q_a(a') \leq U$; or it is $q_{a'}(a') < U$.*

This condition puts a requirement on actions a' that are played with a limit frequency of zero. Roughly, there must be some explanation for why such actions usually have low Q-values. One possibility is that exploration is done only finitely many times. Moreover, if the exploration mechanism is "rigged" in such a way that a' is mostly played when playing policies outside the proximity of $\{\pi_a \mid a \in \text{supp}(p^\Sigma)\}$ then the behaviour of some zero-limit-frequency policies might lead to low Q-values. If exploration of a' is spread out more naturally then all but frequency zero of that exploration will happen near elements of $\{\pi_a \mid a \in \text{supp}(p^\Sigma)\}$, and so a' may not be played with positive frequency if exploring a' near some of the elements of $\{\pi_a \mid a \in \text{supp}(p^\Sigma)\}$ makes a' look poor. If that is not the case then a' will periodically seem optimal, which means that it can only be played with frequency zero if it is quickly learned to be suboptimal when it is played with high probability.

4.2 When is Frequency Convergence Possible?

We believe that there are NDPs in which an ϵ-Greedy agent cannot converge to any limit action frequency. Specifically, let N be the NDP that formalises LARPS as described in Section 3.1, and suppose than an ϵ-Greedy agent has finished some number of episodes in N. Let f_R denote the fraction of past time steps at which a_R was estimated to be the best action, and similarly for f_P and f_S. If we plot the expected direction of change for $\vec{f} = \langle f_R, f_P, f_S \rangle$ we obtain the plot Figure 4. It is presumably not possible to converge to a point that is not an attractor point, and since there is no attractor in this graph, we hence believe that an ϵ-Greedy agent cannot converge to any limit action frequency in this NDP. We have, however, not been able to prove this formally.

We have experimental data to support this argument – Figure 5 depicts five runs of ϵ-Greedy in LARPS. We can see that the agents oscillate between different actions, and that the periods increase in length. This means that there (probably) are cases where not even the action frequencies of an ϵ-Greedy agent converge. Note that this argument applies specifically to tabular ϵ-Greedy agents, and does not necessarily extend to all reinforcement learning agents that cannot converge to mixed policies.

5 Related Work

5.1 Learning Nash Equilibria

If you are playing a symmetric game against an opponent who draws his actions from the same distribution as you then any policy π is ratifiable if and only if $\langle \pi, \pi \rangle$ is a Nash equilibrium (cf. Joyce and Gibbard, 1998, Section 5, on the relationship between Nash equilibria and ratifiability). There is a large body of existing work on learning in the game-theoretic setting. For example, "fictitious play" (due to Brown, 1951) is when some number of players play a

game repeatedly, and in each round play the best response against the empirical distribution over actions taken by their opponent — this is largely analogous to using reinforcement learning. Fudenberg and Levine (1998, Chapter 2) show that fictitious play can only converge to a Nash equilibrium, and that if fictitious play enters a Nash equilibrium it will stay there for all subsequent rounds. It has also been shown that fictitious play can fail to converge (Shapley, 1964). However, there are many special cases in which convergence is guaranteed, including two-player zero-sum games (Robinson, 1951) and generic 2×2 games (Miyasawa, 1961).

5.2 Learning and Newcomblike problems

Other authors have discussed what type of behavior various learning algorithms give rise to in Newcomblike problems. The most common setup is one in which the learner assigns – as she arguably should – values directly to policies, or more generally to that which the agent chooses. It is then usually shown that (among the policies considered) the agent will converge to taking the one with the highest Q-values, or, in decision-theoretical terms, the one with the highest evidential expected utility (Albert and Heiner, 2001; Mayer, Feldmaier, and Shen, 2016; Oesterheld, 2018). This contrasts with our setup, in which the learner selects policies but assigns values to actions. Another more sophisticated setting was studied by Oesterheld (2019a).

6 Discussion and Further Work

We have seen that many of the key assumptions of common reinforcement learning techniques break in the NDP setting. In particular, we have seen that value-based reinforcement learning algorithms can fail to converge to any policy in some NDPs, and that when they do converge, they can only converge to *ratifiable* policies. Philosophers have discussed whether ratifiability should be considered to be a sound normative principle. Note that (as philosophers have pointed out) the policy π that maximises expected discounted reward $\mathbb{E}[R \mid \pi]$ is not in general ratifiable. We have also examined the limit action *frequencies* that agents can converge to (even when the policies do not converge). Still, there are NDPs in which many agents cannot converge even to any such frequency. We gave some results on what limit frequencies are possible. These results are much weaker than our results on policy convergence, because the exact frequencies depend on the details of the learning algorithm. A loose connection to ratifiability can still be drawn.

Overall, established decision-theoretical ideas can be used to understand and formally describe the behavior of "out-of-the-box" reinforcement learning agents in NDPs. However, their behaviour is in general not desirable. They may fail to converge, or they might only be able to converge to suboptimal policies. These algorithms are therefore inappropriate for the Newcomblike setting, which means that there is a need for new kinds of algorithms that explicitly take into account the potential effects of the policy itself.

Throughout the paper, we have noted specific open questions related to our results. For instance, can the results in Section 4.1 be generalised beyond the Bandit setting? There are also many topics and questions about our setting

that we have not touched on at all. For instance, our experimental results indicate that convergence often is slow (considering how simple the given problems are). It might be desirable to back up this impression with theoretical results. We have only studied simple value-based model-free algorithms. Other reinforcement learning algorithms (e.g., policy-gradient or model-based algorithms) may give rise to different considerations. Finally, there are further ways in which we could generalize our setting. One example is to introduce partial observability and imperfect memory into the NDPs. This has been studied in game and decision theory (Piccione and Rubinstein, 1997; Elga, 2000), but recently – under the name *memoryless POMDP* – also in reinforcement learning (Azizzadenesheli, Lazaric, and Anandkumar, 2016; Steckelmacher et al., 2018; cf. Conitzer, 2019). What makes this especially appealing in the NDP context is that problems related to imperfect memory relate closely to Newcomblike problems (Briggs, 2010; Schwarz, 2015). It could also be interesting to more extensively study the properties of discontinuous NDPs.

References

[1] Max Albert and Ronald Asher Heiner. *An Indirect-Evolution Approach to Newcomb's Problem*. CSLE Discussion Paper, No. 2001-01. 2001. URL: https://www.econstor.eu/bitstream/10419/23110/1/2001-01_newc.pdf.

[2] Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. "Open Problem: Approximate Planning of POMDPs in the class of Memoryless Policies". In: *JMLR: Workshop and Conference Proceedings*. Vol. 49. 2016, pp. 1–4.

[3] Richard Bellman. Princeton University Press, 1957.

[4] Steven J. Brams. "Newcomb's Problem and Prisoners' Dilemma". In: *The Journal of Conflict Resolution* 19.4 (Dec. 1975), pp. 596–612.

[5] Rachael Briggs. "Putting a value on Beauty". In: *Oxford Studies in Epistemology*. Vol. 3. Oxford University Press, 2010, pp. 3–24. URL: http://joelvelasco.net/teaching/3865/briggs10-puttingavalueonbeauty.pdf.

[6] Gordon W. Brown. "Iterative Solutions of Games by Fictitious Play". In: *Activity Analysis of Production and Allocation*. Ed. by Tjalling C. Koopmans. John Wiley & Sons and Chapman & Hall, 1951. Chap. XXIV, pp. 371–376. URL: https://archive.org/details/in.ernet.dli.2015.39951/.

[7] Eric G. Cavalcanti. "Causation, Decision Theory, and Bell's Theorem: A Quantum Analogue of the Newcomb Problem". In: *The British Journal for the Philosophy of Science* 61.3 (2010), pp. 569–597. DOI: 10.1093/bjps/axp050.

[8] Vincent Conitzer. "Designing Preferences, Beliefs, and Identities for Artificial Intelligence". In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) Senior Member / Blue Sky Track*. 2019.

[9] Andy Egan. "Some Counterexamples to Causal Decision Theory". In: *Philosophical Review* 116 (Jan. 2007), pp. 93–114. DOI: 10.1215/00318108-2006-023.

[10] Adam Elga. "Self-Locating Belief and the Sleeping Beauty Problem". In: *Analysis* 60.2 (2000), pp. 143–147. URL: https://www.jstor.org/stable/pdf/3329167.pdf.

[11] Drew Fudenberg and David K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[12] David Gauthier. "In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality)". In: *Proceedings of the Aristotelian Society, New Series, 1988–1989*. Vol. 89. 1989, pp. 179–194.

[13] Allan Gibbard and William Harper. "Counterfactuals and Two Kinds of Expected Utility". In: *Foundations and Applications of Decision Theory* 1 (Feb. 1976), pp. 125–162. DOI: 10.1007/978-94-009-9789-9_5.

[14] P. Hall and C.C. Heyde. *Martingale Limit Theory and its Applications*. New York: Academic Press, 1980.

[15] James M. Joyce and Allan Gibbard. "Causal Decision Theory". In: *Handbook of Utility Theory, Volume 1: Principles.* Kluwer, 1998. Chap. 13, pp. 627–666.

[16] Shizuo Kakutani. "A generalization of Brouwer's fixed point theorem". In: *Duke Math. J.* 8.3 (Sept. 1941), pp. 457–459. DOI: 10.1215/S0012-7094-41-00838-4. URL: https://doi.org/10.1215/S0012-7094-41-00838-4.

[17] David Lewis. "Prisoners' Dilemma is a Newcomb Problem". In: *Philosophy & Public Affairs* 8.3 (1979), pp. 235–240.

[18] Dominik Mayer, Johannes Feldmaier, and Hao Shen. "Reinforcement Learning in Conflicting Environments for Autonomous Vehicles". In: *International Workshop on Robotics in the 21st century: Challenges and Promises*. 2016. URL: https://arxiv.org/abs/1610.07089.

[19] Koichi Miyasawa. *On the convergence of the learning process in a 2x2 non-zero-sum two-person game.* Princeton University, Oct. 1961.

[20] Robert Nozick. "Newcomb's Problem and Two Principles of Choice". In: *Essays in Honor of Carl G. Hempel*. Ed. by Nicholas Rescher et al. Springer, 1969, pp. 114–146. URL: http://faculty.arts.ubc.ca/rjohns/nozick_newcomb.pdf.

[21] Caspar Oesterheld. "Approval-directed agency and the decision theory of Newcomb-like problems". In: *Synthese* (2019). ISSN: 1573-0964. DOI: 10.1007/s11229-019-02148-2.

[22] Caspar Oesterheld. *Doing what has worked well in the past leads to evidential decision theory*. 2018. URL: https://casparoesterheld.files.wordpress.com/2018/01/learning-dt.pdf.

[23] Caspar Oesterheld. "Robust Program Equilibrium". In: *Theory and Decision* 86.1 (Feb. 2019), pp. 143–159.

[24] Michele Piccione and Ariel Rubinstein. "On the Interpretation of Decision Problems with Imperfect Recall". In: *Games and Economic Behavior* 20 (1997), pp. 3–24.

[25] Julia Robinson. "An Iterative Method of Solving a Game". In: *Annals of Mathematics* 54.2 (Sept. 1951), pp. 296–301. DOI: 10.2307/1969530.

[26] Gavin Rummery and Mahesan Niranjan. *On-Line Q-Learning Using Connectionist Systems*. Tech. rep. Cambridge University Engineering Department, 1994.

[27] Wolfgang Schwarz. "Lost memories and useless coins: revisiting the absentminded driver". In: *Synthese* 192.9 (2015), pp. 3011–3036.

[28] L. S. Shapley. "Some Topics in Two-Person Games". In: *Advances in Game Theory*. Ed. by M. Dresher, L. S. Shapley, and A. W. Tucker. Princeton University Press, 1964. Chap. 1.

[29] Satinder Singh et al. "Convergence Results for Single-Step On-Policy Reinforcement-Learning Algorithms". In: *Machine Learning* 38 (2000), pp. 287–308. DOI: 10.1023/A:1007678930559.

[30] Denis Steckelmacher et al. "Reinforcement Learning in POMDPs with Memoryless Options and Option-Observation Initiation Sets". In: *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. 2018.

[31] Moshe Tennenholtz. "Program equilibrium". In: *Games and Economic Behavior* 49.2 (2004), pp. 363–373.

[32] H. van Seijen et al. "A theoretical and empirical analysis of Expected Sarsa". In: *2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*. 2009, pp. 177–184.

[33] Chris Watkins. "Learning from delayed rewards". PhD thesis. 1986.

[34] Paul Weirich. "Causal Decision Theory". In: *The Stanford Encyclopedia of Philosophy*. Spring 2016. 2016.

A Q-value convergence

We here show that if a tabular agent converges to a policy π_∞ in a continuous NDP then Q_t converges to q_{π_∞}, assuming that the agent updates its Q-values in an appropriate way. To prove this we will use the following lemma:

Lemma 10. *Let $\langle \zeta_t, \delta_t, F_t \rangle$ be a stochastic process where $\zeta_t, \delta_t, F_t : X \to \mathbb{R}$ satisfy*

$$\delta_{t+1}(x) = (1 - \zeta_t(x_t)) \cdot \delta_t(x_t) + \zeta_t(x_t) \cdot F_t(x_t)$$

with $x_t \in X$ and $t \in \mathbb{N}$. Let P_t be a sequence of increasing σ-fields such that ζ_0 and δ_0 are P_0-measurable and ζ_t, δ_t and F_{t-1} are P_t-measurable, $t \geq 1$. Then δ_t converges to 0 w.p. 1 if the following conditions hold:

1. *X is finite.*
2. *$\zeta_t(x_t) \in [0, 1]$ and $\forall x \neq x_t : \zeta_t(x) = 0$.*
3. *$\sum_t \zeta_t(x_t) = \infty$ and $\sum_t \zeta_t(x_t)^2 < \infty$ w.p. 1.*
4. *$\mathrm{Var}\{F_t(x_t) \mid P_t\} \leq K(1 + \kappa \|\delta_t\|_\infty)^2$ for some $K \in \mathbb{R}$ and $\kappa \in [0, 1)$.*
5. *$\|\mathbb{E}\{F_t \mid P_t\}\|_\infty \leq \kappa \|\delta_t\|_\infty + c_t$, where $c_t \to 0$ w.p. 1 as $t \to \infty$.*

where $\|.\|_\infty$ is a (potentially weighted) maximum norm.

Proof. See Singh et al. (2000). □

We say that a Q-value update rule is *appropriate* if it has the following form;

$$Q_{t+1}(a_t \mid s_t) \leftarrow (1 - \alpha_t(a_t, s_t)) \cdot Q_t(a_t \mid s_t) + \alpha_t(a_t, s_t) \cdot (r_t + \gamma \cdot \hat{v}_{t+1}(s_{t+1})),$$

where $\hat{v}_t(s)$ is an estimate of the value of s, and if moreover

$$\lim_{t \to \infty} \mathbb{E}\left[\hat{v}_t(s) - \max_a Q_t(a \mid s)\right] = 0.$$

Q-learning is of course appropriate. Moreover, SARSA and Expected SARSA are also both appropriate, if the agent is greedy in the limit. Note that since R is bounded, $Q_t(a \mid s)$ has bounded support. This means that if for all $\delta > 0$, $\mathbb{P}(Q_t(\pi_t(s) \mid s) \leq \max_a Q_t(a \mid s) - \delta) \to 0$ as $t \to \infty$, then $\mathbb{E}_{a \sim \pi_t}[Q_t(a \mid s)] \to \max_a Q_t(a \mid s)$ as $t \to \infty$.

Theorem 11. *In any continuous NDP $\langle S, A, T, R, \gamma \rangle$, if a tabular agent converges to a policy π_∞ then Q_t converges to q_{π_∞}, if the following conditions hold:*

1. *The agent updates its Q-values with an appropriate update rule.*
2. *The update rates $\alpha_t(a, s)$ are in $[0, 1)$, and for all $s \in S$ and $a \in A$ we have that $\sum_t \alpha_t(a, s) = \infty$ and $\sum_t \alpha_t(a, s)^2 < \infty$ w.p. 1.*

Note that condition 2 requires that the agent takes every action in every state infinitely many times

Proof. Let

- $X = S \times A$
- $\zeta_t(a, s) = \alpha_t(a, s)$
- $\delta_t(a, s) = Q_t(a \mid s) - q_{\pi_\infty}(a \mid s)$
- $F_t(a, s) = r_t + \gamma \hat{v}_{t+1}(s_{t+1}) - q_{\pi_\infty}(a \mid s)$

Since S and A are finite, and since R is bounded, we have that condition 1 and 4 in Lemma 10 are satisfied. Moreover, assumption 2 of this theorem corresponds to condition 2 and 3 in Lemma 10. It remains to show that condition 5 is satisfied, which we can do algebraically:

$$
\begin{aligned}
&\|\mathbb{E}\{F_t \mid P_t\}\|_\infty \\
&= \max_{s,a} \left| \mathbb{E}\Big[r_t + \gamma \hat{v}_t(s_{t+1}) - q_{\pi_\infty}(a \mid s) \Big] \right| \\
&= \max_{s,a} \Bigg| \mathbb{E}\Big[r_t + \gamma \max_{a'} Q_t(a' \mid s_{t+1}) - q_{\pi_\infty}(a \mid s) + \\
&\qquad \gamma \hat{v}_t(s_{t+1}) - \gamma \max_{a'} Q_t(a' \mid s_{t+1}) \Big] \Bigg| \\
&\leq \max_{s,a} \left| \mathbb{E}\Big[r_t + \gamma \max_{a'} Q_t(a' \mid s_{t+1}) - q_{\pi_\infty}(a \mid s) \Big] \right| + \\
&\quad \max_{s,a} \left| \mathbb{E}\Big[\gamma \hat{v}_t(s_{t+1}) - \gamma \max_{a'} Q_t(a' \mid s_{t+1}) \Big] \right|
\end{aligned}
$$

Note that the second term in this expression is bounded above by

$$
\max_{s} \left| \mathbb{E}\Big[\hat{v}_t(s) - \max_{a} Q_t(a \mid s) \Big] \right|
$$

Let us use k_t to denote this expression. Since the Q-value update rule is appropriate we have that $k_t \to 0$ as $t \to \infty$. We thus have:

$$
= \max_{s,a} \left| \mathbb{E}[r_t + \gamma \max_{a'} Q_t(a' \mid s_{t+1}) - q_{\pi_\infty}(a \mid s)] \right| + k_t
$$

We can now expand the expectations, and rearrange the terms:

$$
\begin{aligned}
&= \max_{s,a} \Bigg| \sum_{s' \in S} \mathbb{P}(T(s,a,\pi_t) \ = s') (\mathbb{E}[R(s,a,s',\pi_t)] \ + \gamma \max_{a'} Q_t(a' \mid s')) \\
&\quad - \ \sum_{s' \in S} \mathbb{P}(T(s,a,\pi_\infty) = s') (\mathbb{E}[R(s,a,s',\pi_\infty)] + \gamma \max_{a'} q_{\pi_\infty}(a' \mid s')) \Bigg| + k_t \\
&= \max_{s,a} \Bigg| \sum_{s' \in S} \mathbb{P}(T(s,a,\pi_\infty) = s') (\mathbb{E}[R(s,a,s',\pi_t)] \ + \gamma \max_{a'} Q_t(a' \mid s') \\
&\qquad\qquad - \mathbb{E}[R(s,a,s',\pi_\infty)] - \gamma \max_{a'} q_{\pi_\infty}(a' \mid s')) \\
&\quad + \ \sum_{s' \in S} (\mathbb{P}(T(s,a,\pi_t) = s') - \mathbb{P}(T(s,a,\pi_\infty) = \ s')) \cdot X \Bigg| + k_t
\end{aligned}
$$

where $X = \mathbb{E}[R(s,a,s',\pi_t)] + \gamma \max_{a'} Q_t(a' \mid s')$. Let $d_t(s,a)$ be the second term in this expression, and let $b_t(s,a,s') = \mathbb{E}\left[R(s,a,s',\pi_t)\right] - \mathbb{E}\left[R(s,a,s',\pi_\infty)\right]$. Since $\pi_t \to \pi_\infty$, and since T and R are continuous, we have that $b_t(s,a,s') \to 0$

and $d_t(s,a) \to 0$ as $t \to \infty$ (for any s, a, and s'). We thus have:

$$
\begin{aligned}
&= \max_{s,a} \Bigg| \sum_{s' \in S} \mathbb{P}(T(s,a,\pi_\infty) = s') \\
&\qquad \Big(\gamma \max_{a'} Q_t(a' \mid s') - \gamma \max_{a'} q_{\pi_\infty}(a' \mid s') + b_t(s,a,s')\Big) + d_t(s,a)\Bigg| + k_t \\
&\leq \gamma \max_{s,a} \Big|Q_t(a \mid s) - q_{\pi_\infty}(a \mid s)\Big| + \max_{s,a,s'} \Big|b_t(s,a,s') + d_t(s,a) + k_t\Big| \\
&= \gamma \max_{s,a} \Big|\delta(s,a)\Big| + c_t = \gamma \|\delta_t\|_\infty + c_t
\end{aligned}
$$

where $c_t = \max_{s,a,s'} \Big|b_t(s,a,s') + d_t(s,a) + k_t\Big|$. This means that

$$\|\mathbb{E}\{F_t \mid P_t\}\|_\infty \leq \gamma \|\delta_t\|_\infty + c_t$$

where $\gamma \in [0,1)$ and $c_t \to 0$ as $t \to \infty$. Thus by lemma 10 we have that Q_t converges to q_{π_∞}. □

B Proof of Theorem 2

Theorem 2. *Let $\mathcal{A}$ be a model-free reinforcement learning agent, and let π_t and Q_t be $\mathcal{A}$'s policy and Q-function at time t. Let $\mathcal{A}$ satisfy the following in a given NDP $\langle S, A, T, R, \gamma \rangle$:*

- *$\mathcal{A}$ is greedy in the limit, i.e. for all $\delta > 0$, $\mathbb{P}(Q_t(\pi_t(s)) \leq \max_a Q_t(a \mid s) - \delta) \to 0$ as $t \to \infty$.*
- *$\mathcal{A}$'s Q-values are accurate in the limit, i.e. if $\pi_t \to \pi_\infty$ then $Q_t \to q_{\pi_\infty}$.*

Then if $\mathcal{A}$'s policy converges to π_∞ then π_∞ is strongly ratifiable on S.

Proof. Let $\pi_t \to \pi_\infty$ and hence $Q_t \to q_{\pi_\infty}$. For strong ratifiability, we have to show that for all actions a' and states s, if a' is suboptimal (in terms of true q values) given π_∞ in s, then $\pi_\infty(a' \mid s) = 0$.

If a' is suboptimal in this way, then there is $\delta > 0$ s.t.

$$q_{\pi_\infty}(a' \mid s) \leq \max_a q_{\pi_\infty}(a \mid s) - \delta.$$

Thus, since $Q_t \to q_{\pi_\infty}$, it is for large enough t,

$$Q_t(a' \mid s) \leq \max_a Q_t(a \mid s) - \frac{\delta}{2}.$$

By the greedy-in-the-limit condition, $\pi_t(a' \mid s) \to 0$. Because $\pi_t \to \pi_\infty$, it follows that $\pi_\infty(a' \mid s) = 0$, as claimed. □

C Proof of Theorem 3

Lemma 12 (Kakutani's Fixed-Point Theorem)**.** *Let X be a non-empty, compact, and convex subset of some Euclidean space $\mathbb{R}^n$, and let $\phi : X \to 2^X$ be a set-valued function s.t. ϕ has a closed graph and s.t. $\phi(x)$ is non-empty and convex for all $x \in X$. Then ϕ has a fixed point.*

Proof. See Kakutani (1941). □

Theorem 3. *Every continuous NDP has a strongly ratifiable policy.*

Proof. Let $N = \langle S, A, T_N, R_N, \gamma \rangle$ be a continuous NDP, and let N_π be the MDP $\langle S, A, T_{N_\pi}, R_{N_\pi}, \gamma \rangle$ that is obtained by fixing the dynamics in N according to π – that is, $T_{N_\pi}(s, a) = T_N(s, a, \pi)$, and $R_{N_\pi}(s, a, s') = R_N(s, a, s', \pi)$. Let $\phi_N : (S \rightsquigarrow A) \to 2^{(S \rightsquigarrow A)}$ be the set-valued function s.t. $\phi_N(\pi)$ is the set of all policies that are optimal in N_π. We will show that the graph of ϕ_N is closed and apply Kakutani's fixed point theorem.

Suppose (π_i) is a sequence of policies converging to π_0 and suppose $\lambda_i \in \phi_N(\pi_i)$ is a sequence converging to λ_0. For all sufficiently large i, $\text{supp}(\lambda_0) \subseteq \text{supp}(\lambda_i)$ (as the state and action spaces are finite). Therefore for sufficiently large i, $\lambda_0 \in \phi_N(\pi_i)$. By the continuity with respect to π of $\mathbb{E}[R \mid \lambda_0]$ in N_π, $\lambda_0 \in \phi_N(\pi_0)$. Therefore, the graph of ϕ_N is closed.

The domain of ϕ_N is a non-empty, compact, convex subset of Euclidean space. Any MDP always has an optimal policy, and so $\phi_N(\cdot)$ is non-empty. Since N_π is an MDP $\phi_N(\pi)$ is a set of deterministic policies and all their convex combinations, and so $\phi_N(\cdot)$ is convex. Hence, by Kakutani's Fixed Point Theorem, there must be a π s.t. $\pi \in \phi_N(\pi)$. Then π is strongly ratifiable in N. Hence every continuous NDP has a strongly ratifiable policy. □

D Proof of Theorem 6

To prove Theorem 6, we first need to prove the following lemma.

Lemma 13. *Let X_t be a non-negative discrete stochastic process, indexed by t, and let $\mathcal{F}_t$ denote the history upto time t. Suppose X_t is bounded, i.e. there exists B such that $X_t \leq B$, and further that $|X_{t+1} - X_t| < B/t$. Suppose also that there exists $\epsilon > 0$ and $b > 0$ such that whenever $X_t < b$,*

$$Var(X_{t+1}|\mathcal{F}_t) \geq \frac{\epsilon}{t^2} \tag{5}$$

and

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] - X_t \geq 0. \tag{6}$$

Then $\mathbb{P}(X_t \to 0) = 0$.

Proof. Let $a_n = 2^{2^n}$ and define the following sequences of events. Firstly, letting s_n denote $2^n \sqrt{4B^2 \sum_{t=a_{n+1}}^{\infty} \frac{1}{t^2}}$,

$$A_n = \left\{ X_{a_{n+1}} > s_n \right\} \tag{7}$$

and

$$A'_n = A_n \vee \left\{ \exists t \in [a_n, a_{n+1}] \text{ s.t. } X_t \geq b \right\}, \tag{8}$$

which tell us that at some point after time a_n, but not after a_{n+1}, the value of X_t isn't very small and secondly

$$B_n = \left\{ X_t < b \forall t \geq a_n \right\}. \tag{9}$$

This event is useful because it is implied by convergence to 0 and tells us that Equation 6 can be applied.

We will show that two properties hold. Firstly that $\mathbb{P}(A'_n \wedge B_n \wedge \{X_t \to 0\}) \leq 2^{-2n}$ and secondly that $\mathbb{P}(A'_n|\mathcal{F}_{a_n}) \geq 2/5$ for all sufficiently large n.

From the second of these properties, and the fact that A'_n is $\mathcal{F}_{a_{n+1}}$ measurable, it is immediate by the argument of the Borel-Cantelli Lemma that, almost surely, A'_n occurs infinitely often (i.o.) i.e. for infinitely many n. From this and the fact that $X_t \to 0 \implies (B_n \forall n$ sufficiently large) we can deduce the following

$$\begin{aligned}
\mathbb{P}(X_t \to 0) &= \mathbb{P}(B_n \wedge \{X_t \to 0\} \forall n \text{ sufficiently large}) && (10)\\
&= \mathbb{P}((A'_n \wedge B_n \wedge \{X_t \to 0\}) \text{ i.o.}) && (11)\\
&\leq \mathbb{P}(\exists n > m \text{ s.t. } A'_n \wedge B_n \wedge \{X_t \to 0\}) && (12)\\
&\leq \sum_{n=m}^{\infty} \mathbb{P}(A'_n \wedge B_n \wedge \{X_t \to 0\}). && (13)
\end{aligned}$$

It is immediate from the first fact that this sum is convergent, and thus it must converge to zero as $m \to \infty$, but m was arbitrary so $\mathbb{P}(X_t \to 0) = 0$.

We now prove the first property. Note that if B_n occurs then A'_n can only occur if A_n occurs. Thus $\mathbb{P}(A'_n \wedge B_n \wedge \{X_t \to 0\}) \leq \mathbb{P}(B_n \wedge \{X_t \to 0\}|A_n)$. To see this is small, we consider an augmentation of X_t given by

$$Y_t = \begin{cases} X_t & t \leq a_{n+1} \\ Y_{t-1} + (X_t - X_{t-1}) - \mathbb{E}[X_t - X_{t-1}] & t > a_{n+1}. \end{cases} \tag{14}$$

Note that this process is a martingale (for $t > a_{n+1}$), i.e. $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] = Y_t$ for all $t > a_{n+1}$, and that if B_n occurs then $Y_t \leq X_t$ for all t (by Equation 6). As Y is a martingale $\mathbb{E}[Y_t|\mathcal{F}_{a_{n+1}}] = Y_{a_{n+1}}$. Furthermore we can compute as follows

$$\begin{aligned}
\mathrm{Var}(Y_t|\mathcal{F}_{a_{n+1}}) &= \mathbb{E}[(Y_t - Y_{a_{n+1}})^2|\mathcal{F}_{a_{n+1}}] && (15)\\
&= \mathbb{E}[(\sum_{r=a_{n+1}}^{t-1} Y_{r+1} - Y_r)^2|\mathcal{F}_{a_{n+1}}] && (16)\\
&= \mathbb{E}[\sum_{r=a_{n+1}}^{t-1} \sum_{s=a_{n+1}}^{t-1} (Y_{r+1} - Y_r)(Y_{s+1} - Y_s)|\mathcal{F}_{a_{n+1}}] && (17)\\
&= \sum_{r=a_{n+1}}^{t-1} \sum_{s=a_{n+1}}^{t-1} \mathbb{E}[(Y_{r+1} - Y_r)(Y_{s+1} - Y_s)|\mathcal{F}_{a_{n+1}}]. && (18)
\end{aligned}$$

As Y is a martingale we have that this final expectation is zero unless $r = s$. To see this assume WLOG that $r > s$ and note that

$$\begin{aligned}
\mathbb{E}[(Y_{r+1} - Y_r)(Y_{s+1} - Y_s)|\mathcal{F}_{a_{n+1}}] &= \mathbb{E}[\mathbb{E}[(Y_{r+1} - Y_r)(Y_{s+1} - Y_s)|\mathcal{F}_r]|\mathcal{F}_{a_{n+1}}] && (19)\\
&= \mathbb{E}[\mathbb{E}[(Y_{r+1} - Y_r)|\mathcal{F}_r)(Y_{s+1} - Y_s)|\mathcal{F}_{a_{n+1}}] && (20)\\
&= \mathbb{E}[0(Y_{s+1} - Y_s)|\mathcal{F}_{a_{n+1}}] && (21)\\
&= 0. && (22)
\end{aligned}$$

Putting these together, along with the fact that $Y_{r+1} - Y_r \leq 2B/r$ (which follows from the similar bound on difference in X), we get that

$$\mathrm{Var}(Y_t|\mathcal{F}_{a_{n+1}}) = \sum_{r=a_{n+1}}^{t-1} \mathbb{E}[(Y_{r+1} - Y_r)^2|\mathcal{F}_{a_{n+1}}] \tag{23}$$

$$\leq 4B^2 \sum_{r=a_{n+1}}^{\infty} r^{-2}. \tag{24}$$

Thus, for all $t \geq a_{n+1}$, by Chebyshev's inequality,

$$\mathbb{P}(Y_t < 0|A_n) \leq \mathbb{P}(|Y_t - Y_{a_{n+1}}| > Y_{a_{n+1}}|A_n) \tag{25}$$

$$\leq \mathbb{P}\left(|Y_t - Y_{a_{n+1}}| > s_n|A_n\right) \tag{26}$$

$$\leq \frac{\mathrm{Var}(Y_t|\mathcal{F}_{a_{n+1}})}{s_n^2} \tag{27}$$

$$\leq 2^{-2n}. \tag{28}$$

Whilst by the final property if B_n occurs and $X_t \to 0$ then $Y_t < \eta$ for all sufficiently large t for all $\eta > 0$. Thus $\mathbb{P}(B_n \wedge \{X_t \to 0\}|A_n) \leq 2^{-2n}$ and $\mathbb{P}(A'_n \wedge B_n \wedge \{X_t \to 0\}) \leq 2^{-2n}$.

We now prove that $\mathbb{P}(A'_{n+1}|\mathcal{F}_{a_{n+1}}) \geq 2/5$ for sufficiently large n, where we have replaced n by $n+1$ for convenience. We again define Y_t exactly as for the previous property and note again that it is a martingale and that, for $t \geq a_{n+1}$, $4B^2/t^2 \geq \mathrm{Var}(Y_{t+1}|\mathcal{F}_t) \geq \epsilon/t^2$. Thus we can apply the martingale central limit theorem (Hall and Heyde, 1980, Theorem 5.4) to conclude that, setting $\sigma_n^2 = \mathrm{Var}(Y_{a_{n+1}} - Y_{a_n}|\mathcal{F}_{a_n})$, the distribution conditioned on $\mathcal{F}_{a_{n+1}}$ of $(Y_{a_{n+2}} - Y_{a_{n+1}})/\sigma_{n+1}$ converges to a standard normal distribution as $n \to \infty$. Let Z have a standard normal distribution.

$$\begin{aligned}
&\mathbb{P}(Y_{a_{n+2}} > s_{n+1}) \\
=&\mathbb{P}((Y_{a_{n+2}} - Y_{a_{n+1}})/\sigma_{n+1} > (s_{n+1} - Y_{a_{n+1}})/\sigma_{n+1}) \\
=&\mathbb{P}((Y_{a_{n+2}} - Y_{a_{n+1}})/\sigma_{n+1} > (s_{n+1} - X_{a_{n+1}})/\sigma_{n+1}) \\
\geq&\mathbb{P}((Y_{a_{n+2}} - Y_{a_{n+1}})/\sigma_{n+1} > s_{n+1}/\sigma_{n+1}) \\
\to&\mathbb{P}(Z > \lim_{n\to\infty} s_{n+1}/\sigma_{n+1}) \\
=&\mathbb{P}(Z > 0) = \frac{1}{2}
\end{aligned}$$

Where the limit in the probability was zero because $s_{n+1} = O(2^{n+1-3\cdot 2^{n+1}})$ and $\sigma_{n+1} = \Omega(2^{-3\cdot 2^n})$. Finally note that, $X_t \geq Y_t$ for all $t \leq a_{n+2}$ unless the event $\{\exists a_{n+1} \leq t \leq a_{n+2} \mathrm{s.t.} X_t \geq b\}$ occurs. So for sufficiently large n either $\{\exists a_{n+1} \leq t \leq a_{n+2} \mathrm{s.t.} X_t \geq b\}$ or, with probability at least $2/5$, A_{n+1} occurs. Therefore, for sufficiently large n, $\mathbb{P}(A'_{n+1}|\mathcal{F}_{a_{n+1}}) \geq 2/5$ and the proof is complete. □

Theorem 6. *Let $\mathcal{A}$ be an agent that plays the Repellor Problem, explores infinitely often, and updates its Q-values with a learning rate α_t that is constant across actions, and let π_t and Q_t be $\mathcal{A}$'s policy and Q-function at time t. Assume also that for $j \neq i$, if $\pi_t(a_i)$, $\pi_t(a_j)$ both converge to positive values, then*

$$\frac{\pi_t(a_i) - \pi_t(a_j)}{Q_t(a_i) - Q_t(a_j)} \underset{a.s.}{\to} \infty \tag{2}$$

as $t \to \infty$. Then π_t almost surely does not converge as $t \to \infty$.

Proof. We first need to establish the fact that $(1/3, 1/3, 1/3)$ is the only strongly ratifiable policy. First, if $\pi(a_j) \leq 1/4$ for some j then $\mathbb{E}[R(a_i, \pi)] = \pi(a_{i+1})$. It is easy to see that for this reward function, there is no strongly ratifiable policy other than the symmetric $(1/3, 1/3, 1/3)$.

The other case of $\pi(a_j) > 1/4$ for all j is harder. Finding strongly ratifiable policies in this range gives rise to the following system of polynomial equations, constrained to $p_1, p_2, p_3 \in [1/4, 1]$:

$$\begin{aligned} p_1 + 4 \cdot 13^3 p_2 \left(p_1 - \frac{1}{4}\right)\left(p_2 - \frac{1}{4}\right)\left(p_3 - \frac{1}{4}\right) &= x \\ p_2 + 4 \cdot 13^3 p_3 \left(p_1 - \frac{1}{4}\right)\left(p_2 - \frac{1}{4}\right)\left(p_3 - \frac{1}{4}\right) &= x \\ p_3 + 4 \cdot 13^3 p_1 \left(p_1 - \frac{1}{4}\right)\left(p_2 - \frac{1}{4}\right)\left(p_3 - \frac{1}{4}\right) &= x \\ p_1 + p_2 + p_3 &= 1 \end{aligned}$$

Although this is non-trivial, it can be solved by computer algebra system.[3] For completeness, we would like to give a more human argument here. Consider the simpler system

$$p_1 + Kp_2 = p_2 + Kp_3 = p_3 + Kp_1 \tag{29}$$

$$p_1 + p_2 + p_3 = 1 \tag{30}$$

Note that for p_1, p_2, p_3 to satisfy the original system of equations, it has to satisfy the above system of equations for a particular $K > 0$. It turns out that even without knowing K, the unique solution to this equation system is the symmetric $p_1 = p_2 = p_3$. To prove this, assume that the three are not the same. WLOG we can assume that p_1 is among the maxima of $\{p_1, p_2, p_3\}$. Then we can distinguish two cases: First, imagine that $p_1 \geq p_2 \geq p_3$, where at least one of the two inequalities is strict. Then because $K > 0$, it is $p_1 + Kp_2 > p_2 + Kp_3$, contradicting the first equality in line 29. Second, imagine that $p_1 \geq p_3 \geq p_2$, where at least one of the inequalities is strict. Then it is $p_2 + Kp_3 < p_3 + Kp_1$, contradicting the second equality in line 29. In conclusion, it must be $p_1 = p_2 = p_3$ as claimed.

Now that we have shown that $(1/3, 1/3, 1/3)$ is the only strongly ratifiable policy, we can conclude by Theorem 2, that π_t almost surely does not converge to any policy other than $(1/3, 1/3, 1/3)$. It now only remains to show that π_t almost surely does not converge to $(1/3, 1/3, 1/3)$.

To show that π_t cannot converge to $(1/3, 1/3, 1/3)$, we will analyze the history of what we will call *relative (empirical) Q-values*, which we will denote by $D_t(a_j, a_i) = Q_t(a_j) - Q_t(a_i)$. In order to converge to $(1/3, 1/3, 1/3)$, the relative Q-values must all converge to 0. In particular, it has to be

$$X_t \coloneqq \sum_{a_i, a_j : i < j} |D_t(a_j, a_i)| \to 0, \tag{31}$$

[3]For example, in Mathematica, the following code identifies the unique solution $(1/3, 1/3, 1/3)$: `Solve[(4*13^3) * p1 * ((p1-1/4)*(p2-1/4)*(p3-1/4)) + p2 == (4*13^3) * p2 * ((p1-1/4)*(p2-1/4)*(p3-1/4)) +p3 == (4*13^3) * p3 * ((p1-1/4) * (p2-1/4)*(p3-1/4)) + p1 && p1+p2+p3==1 && p1>=1/4 && p2>=1/4 && p3>=1/4, p1,p2,p3]`

as $t \to \infty$.

We will show, however, that these values almost surely do not converge to 0 if the policies converge to $(1/3, 1/3, 1/3)$. Roughly, we show that when the relative Q-values are close to 0 and the agent acts according to a policy that is close to $(1/3, 1/3, 1/3)$, the Q-values will in expectation be updated toward the action that is currently most likely to be taken. Thus for large enough t, X_t will always increase in expectation. With some other easy-to-verify properties of X_t, we can then apply Lemma 13, which gives us that almost surely the X_t do not converge to 0 as $t \to \infty$.

In order to prove that $\mathbb{E}\left[X_t \mid \mathcal{F}_{t-1}\right] - X_{t-1} > 0$ for large enough t and assuming X_t is close to 0 and π_t close to $(1/3, 1/3, 1/3)$, let $a^* \in \arg\max_a \pi_t(a)$. Because of stochasticity of the rewards and by line 2, it is $\pi_t(a^*) > 1/3$ for large enough t. Further, let $a^- \in \arg\min_a \pi_t(a)$. It is $\pi_t(a^-) \leq 1/3$. Finally, let $\epsilon = \pi_t(a^*) - \pi_t(a^-)$.

The $X_t - X_{t-1}$ can be seen as the sum of three differences $|D_t(a_j, a_i)| - |D_{t-1}(a_j, a_i)|$. We start with the difference for a^* and a^-. It is

$$\begin{aligned} &\mathbb{E}\left[|D_t(a^*, a^-)| \mid \mathcal{F}_{t-1}\right] - |D_{t-1}(a^*, a^-)| \\ &\quad = \alpha_t \left(\mathbb{E}\left[R(a^*, \pi_t)\right] - \mathbb{E}\left[R(a^-, \pi_t)\right]\right) - \alpha_t \left(Q_{t-1}(a^*) - Q_{t-1}(a^-)\right) \end{aligned} \tag{32}$$

Now, assuming that π is close enough to $(1/3, 1/3, 1/3)$ that $\pi(a_j) \geq 1/4 + 1/13$ for all j, it is

$$\begin{aligned} & \mathbb{E}\left[R(a^*, \pi_t)\right] - \mathbb{E}\left[R(a^-, \pi_t)\right] && (33) \\ = \;& \pi(a^*_{+1}) - \pi(a^-_{+1}) + (\pi(a^*) - \pi(a^-)) \cdot 4 \cdot 13^3 \prod_j \left(\pi(a_j) - \frac{1}{4}\right) && (34) \\ \geq \;& -\epsilon + 4\epsilon && (35) \end{aligned}$$

It is left to estimate the other summands in the expectation of $X_t - X_{t-1}$. Consider any pair of actions a_i, a_j with $i > j$. Because $|D_t(a_i, a_j)| = |D_t(a_j, a_i)|$, we can assume WLOG that $Q_{t-1}(a_i) > Q_{t-1}(a_j)$, which for large enough t also means $\pi_t(a_i) > \pi_t(a_j)$. Thus, by similar reasoning as before,

$$\begin{aligned} &\mathbb{E}\left[|D_t(a_i, a_j)| \mid \mathcal{F}_{t-1}\right] - |D_{t-1}(a_i, a_j)| \\ &\quad = \alpha_t \left(\mathbb{E}\left[R(a_i, \pi_t)\right] - \mathbb{E}\left[R(a_j, \pi_t)\right]\right) - \alpha_t \left(Q_{t-1}(a_i) - Q_{t-1}(a_j)\right). \end{aligned} \tag{36}$$

and

$$\mathbb{E}\left[R(a_i, \pi_t)\right] - \mathbb{E}\left[R(a_j, \pi_t)\right] \geq -\epsilon. \tag{37}$$

Thus, overall for large enough t we have

$$\begin{aligned} & \mathbb{E}\left[X_t \mid \mathcal{F}_t\right] - X_{t-1} && (38) \\ \geq \;& \alpha_t \epsilon - \alpha_t \left(\sum_{a_i, a_j : i < j} Q_{t-1}(a_i) - Q_{t-1}(a_j) \right) && (39) \end{aligned}$$

By line 2, ϵ outgrows the differences in Q-values and therefore this term will be positive for all large enough t, as claimed. □

E Proof of Theorem 7

Theorem 7. *Assume that there is some sequence of random variables* $(\epsilon_t \geq 0)_t$ *s.t.* $\epsilon_t \underset{t\to\infty}{\underset{a.s.}{\to}} 0$ *and for all* $t \in \mathbb{N}$ *it is*

$$\sum_{a^* \in \arg\max_a Q_t(a)} \pi_t(a^*) \geq 1 - \epsilon_t. \tag{4}$$

Let $P_t^\Sigma \to p^\Sigma$ *with positive probability as* $t \to \infty$*. Then across all actions* $a \in \text{supp}(p^\Sigma)$*,* $q_a(a)$ *is constant.*

Proof. Consider any $a \in \text{supp}(p^\Sigma)$ that is played with positive frequency. Because exploration goes to zero, almost all (i.e. frequency 1) of the time that a is played must be from π_t playing a with probability close to 1. Therefore, whenever $P_t^\Sigma \underset{t\to\infty}{\to} p^\Sigma$ it is

$$Q_t(a) \underset{t\to\infty}{\underset{\text{a.s.}}{\to}} q_a(a). \tag{40}$$

Thus $q_a(a)$ must be constant across $a \in \text{supp}(p^\Sigma)$, since otherwise the actions with lower values of $q_a(a)$ could not be taken in the limit. □

F Proof of Theorem 8

Theorem 8. *Same assumptions as Theorem 7. If* $|\text{supp}(p^\Sigma)| > 1$ *then for all* $a \in \text{supp}(p^\Sigma)$ *there exists* $a' \in A$ *s.t.* $q_a(a') \geq q_a(a)$*.*

Proof. Let $|\text{supp}(p^\Sigma)| > 1$ and suppose that $\exists a \in \text{supp}(p^\Sigma)$ s.t.

$$\forall a' \in A - \{a\} \colon q_a(a') < q_a(a). \tag{41}$$

Policies close to π_a are almost surely played infinitely often. Every time T this happens we have that $Q_T(a) \geq Q_T(a')$ for all $a' \in A - \{a\}$. Now it is easy to see that if 41 holds, then there is a K s.t. every such time T, there is a chance of at least K that for all $t \geq T$ it is $Q_t(a) > Q_t(a')$ for all $a' \in A - \{a\}$. Hence almost surely $\text{supp}(p^\Sigma) = \{a\}$, which contradicts the assumption that $|\text{supp}(p^\Sigma)| > 1$. □

G Proof of Theorem 9

Theorem 9. *Same assumptions as Theorem 7. Let* U *be the Q-value* $q_a(a)$ *which (by Theorem 7) is constant across* $a \in \text{supp}(p^\Sigma)$*. For any* $a' \in A - \text{supp}(p^\Sigma)$ *that is played infinitely often, let frequency 1 of the exploratory plays of* a' *happen when playing a policy near elements of* $\{\pi_a \mid a \in \text{supp}(p^\Sigma)\}$*. Then there is* $a \in \text{supp}(p^\Sigma)$ *such that* $q_a(a') \leq U$*; or it is* $q_{a'}(a') < U$*.*

Proof. Suppose there is an $a' \in A - \text{supp}(p^\Sigma)$ for which both are false, i.e. $q_a(a') > U$ for all $a \in \text{supp}(p^\Sigma)$, and $q_{a'}(a') \geq U$. Frequency 1 of the time that a' is played is when the policy is near an element of $\{\pi_a \mid a \in \text{supp}(p^\Sigma) \cup \{a'\}\}$, and so $Q_t(a')$ converges to some convex combination of $q_a(a')$ for $a \in \text{supp}(p^\Sigma) \cup \{a'\}$. Therefore, in the limit $Q_t(a')$ is bigger than U. But that is inconsistent with a' being played with frequency 0. □

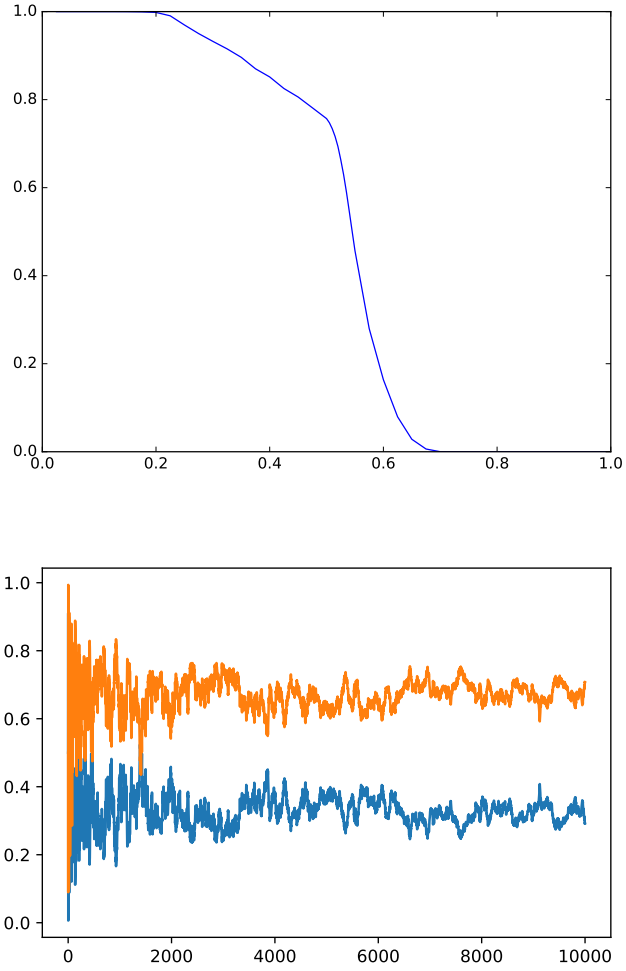

Figure 2: The upper figure plots the probability of Softmax converging to the mixed strategy in Death in Damascus given $\beta_n = n^{-\alpha}$ against α. More accurately it is a plot of the fraction of runs which assigned a Q-value of at least 5.5 to the action of going to Aleppo after 5000 iterations. These are empirical probabilities from 20,000 runs for every α that is a multiple of 0.025 and 510,000 for each α that is a multiple of 0.005 between 0.5 and 0.55. Notice the "kink" at $\alpha = 0.5$. Based on our experiments, this kink is not an artefact and shows up reliably in this kind of graph. The bottom figure shows how the action probabilities evolve over time for a single run (chosen to converge to the mixed strategy) for $\alpha = 0.3$.

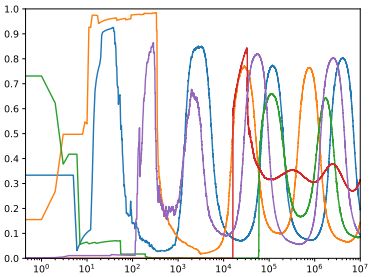

Figure 3: This figure shows five runs of a softmax agent in LARPS, and plots $\pi(a_{\text{rock}})$ against the total number of episodes played. The agent's Q-values are the historical mean rewards for each action, and $\beta_t = 1/\log t$.

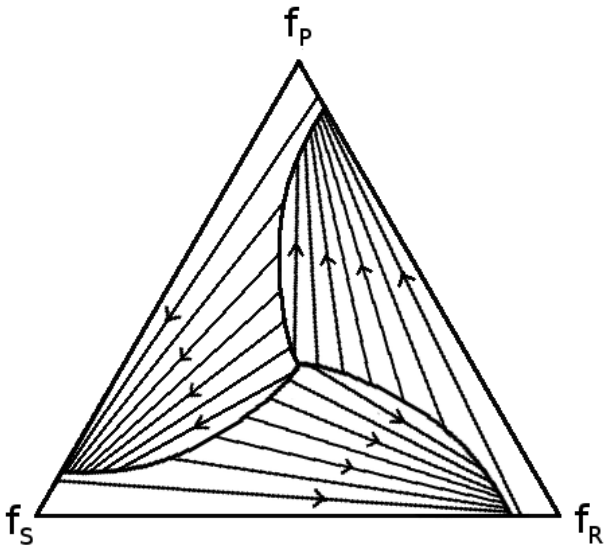

Figure 4: The dynamics of LARPS.

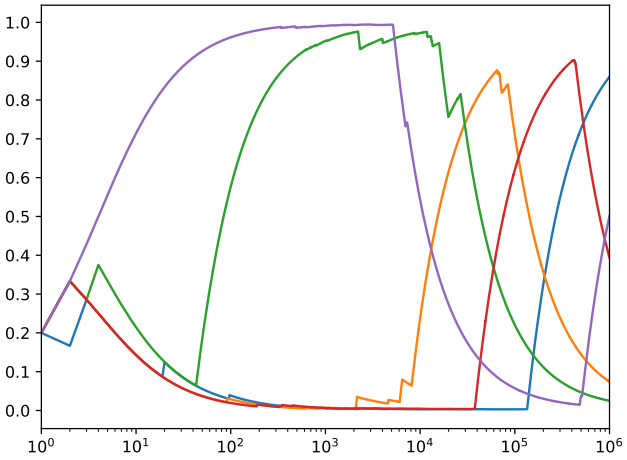

Figure 5: This figure shows five runs of an ϵ-Greedy agent in LARPS, and plots the proportion of past episodes in which the agent played "rock" against the total number of episodes played. The agent's Q-values are the historical mean rewards for each action, and its ϵ-value is 0.01.