Environments cannot detect Markov agents

Samuel Allen Alexander[1 [0000−0002−7930−110X]]

The U.S. Securities and Exchange Commission samuelallenalexander@gmail.com
https://philpeople.org/profiles/samuel-alexander/publications

Abstract. We introduce a method for combining two reinforcement learning agents (equivalently: two decision theoretical agents) into a new agent with the property that the expected total reward the new agent gets in each environment is the average of the expected total rewards the two original agents get in that environment. Using this, we formalize and strengthen an informal result of Alexander and Hutter, and we prove a surprising additional result. Call an agent "Markov" if that agent ignores all training data. We show that no environment can reward agents for being Markov. More precisely: if an environment gives positive expected total reward to every Markov agent, then it must give positive expected total reward to some non-Markov agent. We informally argue that this result casts doubt on an informal conjecture of Silver et al.

1 Introduction

In reinforcement learning, an agent π interacts with an environment μ. The agent and the environment take turns.

- On π's turn, π outputs a probability distribution over a fixed action-set. Based on this distribution, an action is randomly chosen and is transmitted to π and μ.
- On μ's turn, μ outputs a probability distribution over a fixed percept-set, where every percept includes an observation (thought of as the agent's view of the world) and a numerical reward. Based on this distribution, a percept is randomly chosen and is transmitted to π and μ.

These turns continue forever, and the whole sequence of turns is called an agent-environment interaction.

If π and ρ are two agents, we can informally imagine a new agent σ (first described in [AH21]) as follows. At the beginning of every agent-environment interaction, σ flips a coin. If the coin lands heads, then σ transforms into π; otherwise, σ transforms into ρ. Note that the coin is only flipped one time, at the very start of the agent-environment interaction: it is not repeatedly flipped every turn. Intuitively, it seems like the expected total reward in the agent-environment interaction when σ interacts with μ, should be the average of the corresponding expected total rewards when π or ρ interact with μ. But this is all quite informal, as the reinforcement learning framework does not actually provide any mechanism for such an initial coin-flip.

We will show that there is a way to combine π and ρ into a new agent $\pi \oplus \rho$, within the formal framework, such that the expected total reward when $\pi \oplus \rho$ interacts with μ is the average of the expected total rewards when π or ρ interact with μ. Thus, at least up to expected total reward, $\pi \oplus \rho$ has the exact same performance as the above informal coin-flipping agent.

The structure of this paper is as follows.

...

2 Preliminaries

(Define RL)

3 The $\oplus$ operator

Definition 1. *Suppose s is a finite sequence of alternating percepts and actions. Let π be an agent. Let μ be an environment.*

1. *The* probability of s according to π, *written $\pi_*(s)$, is the probability that when π and μ interact, the interaction begins with s, given that all percepts are as in s. Formally we define $\pi_*(s)$ by induction:*
   (a) *If s contains no action, then $\pi_*(s) = 1$.*
   (b) *If $s = t \frown a$ (where a is an action) or $s = t \frown a \frown p$ (where a is an action and p is a percept), then $\pi_*(s) = \pi_*(t)\pi(a|t)$.*
2. *The* probability of s according to μ, *written $\mu_*(s)$, is the probability that when π and μ interact, the interaction begins with s, given that all actions are as in s. Formally we define $\mu_*(s)$ by induction:*
   (a) *If s has length 0, then $\mu_*(s) = 1$.*
   (b) *If $s = t \frown p$ (where p is a percept) or $s = t \frown p \frown a$ (where a is an action and p is a percept), then $\mu_*(s) = \mu_*(t)\mu(p|t)$.*
3. *The* probability of s according to π and μ, *written $(\pi, \mu)_*(s)$, is the probability that when π and μ interact, the interactin begins with s. Formally, we define $(\pi, \mu)_*(s)$ by induction:*
   (a) *If s has length 0, then $(\pi, \mu)_*(s) = 1$.*
   (b) *If $s = t \frown p$ (where p is a percept), then $(\pi, \mu)_*(s) = (\pi, \mu)_*(t)\mu(p|t)$.*
   (c) *If $s = t \frown a$ (where a is an action), then $(\pi, \mu)_*(s) = (\pi, \mu)_*(t)\pi(a|t)$.*

Lemma 1. *For all s, π, μ as in Definition 1,*

$$(\pi, \mu)_*(s) = \pi_*(s)\mu_*(s).$$

Proof. By induction.

Lemma 2. $V^\pi_{\mu,n} = \sum_{s \in S_n} (\pi, \mu)_*(s) r(s)$, *where S_n is the set of all length-n initial percept-action sequences and $r(s)$ is the total reward in s.*

Proof. Basic probability theory.

Definition 2. *Assume π and ρ are agents. Define the new agent $\pi \oplus \rho$ by*

$$(\pi \oplus \rho)(a|s) = \frac{\pi_*(s \frown a) + \rho_*(s \frown a)}{\pi_*(s) + \rho_*(s)}$$

provided $\pi_*(s) + \rho_*(s) > 0$; otherwise, let $(\pi \oplus \rho)(a|s) = 1/|\mathcal{A}|$.

Lemma 3. *If π and ρ are agents then $\pi \oplus \rho$ really is an agent.*

Proof. We must show $\sum_{a \in \mathcal{A}} (\pi \oplus \rho)(a|s) = 1$ for every percept-action sequence s ending in a percept. Fix some such s.

Case 1: $\pi_*(s) = \rho_*(s) = 0$. Then by definition each $(\pi \oplus \rho)(a|s) = 1/|\mathcal{A}|$ so the claim is immediate.

Case 2: $\pi_*(s) + \rho_*(s) > 0$. Then

$$\begin{aligned}
\sum_{a \in \mathcal{A}} (\pi \oplus \rho)(a|s) &= \sum_{a \in \mathcal{A}} \frac{\pi_*(s \frown a) + \rho_*(s \frown a)}{\pi_*(s) + \rho_*(s)} && \text{(Definition 2)} \\
&= \sum_{a \in \mathcal{A}} \frac{\pi_*(s)\pi(a|s) + \rho_*(s)\rho(a|s)}{\pi_*(s) + \rho_*(s)} && \text{(Definition 1)} \\
&= \frac{\pi_*(s)\left(\sum_{a \in \mathcal{A}} \pi(a|s)\right) + \rho_*(s)\left(\sum_{a \in \mathcal{A}} \rho(a|s)\right)}{\pi_*(s) + \rho_*(s)} && \text{(Algebra)} \\
&= \frac{\pi_*(s) + \rho_*(s)}{\pi_*(s) + \rho_*(s)} = 1. && (\pi, \rho \text{ are agents})
\end{aligned}$$

Theorem 1. *For all agents π and ρ, for every environment μ, if V_μ^π and V_μ^ρ converge, then*

$$V_\mu^{\pi \oplus \rho} = \frac{1}{2}(V_\mu^\pi + V_\mu^\rho).$$

Proof. It suffices to show that for every $n \in \mathbb{N}$, $V_{\mu,n}^{\pi \oplus \rho} = \frac{1}{2}(V_{\mu,n}^\pi + V_{\mu,n}^\rho)$. Let $n \in \mathbb{N}$. Let S_n be the set of all length-n percept-action sequences, and for each $s \in S_n$, let $r(s)$ be the sum of rewards in s. By Lemma 2, it suffices to show that $(\pi \oplus \rho, \mu)_*(s) = \frac{1}{2}((\pi, \mu)_*(s) + (\rho, \mu)_*(s))$ for each $s \in S_n$. Fix $s \in S_n$.

Case 1: $s = \langle\rangle$. Then $(\pi \oplus \rho, \mu)_*(s) = 1 = \frac{1}{2}(1+1) = \frac{1}{2}((\pi, \mu)_*(s) + (\rho, \mu)_*(s))$.

Case 2: $s = t \frown p$ for some percept p. Then

$$\begin{aligned}
(\pi \oplus \rho, \mu)_*(s) &= (\pi \oplus \rho, \mu)_*(t \frown p) \\
&= (\pi \oplus \rho, \mu)_*(t)\mu(p|t) && \text{(Definition 1)} \\
&= ((\pi, \mu)_*(t) + (\rho, \mu)_*(t))\mu(p|t)/2 && \text{(Induction)} \\
&= ((\pi, \mu)_*(t)\mu(p|t) + (\rho, \mu)_*(t)\mu(p|t))/2 && \text{(Algebra)} \\
&= ((\pi, \mu)_*(t \frown p) + (\rho, \mu)_*(t \frown p))/2 && \text{(Definition 1)} \\
&= ((\pi, \mu)_*(s) + (\rho, \mu)_*(s))/2,
\end{aligned}$$

as desired.

Case 3: $s = t \frown a$ for some action a.

Subcase 3.1: $\pi_*(t) = \rho_*(t) = 0$. By Lemma 1, $(\pi,\mu)_*(t) = (\rho,\mu)_*(t) = 0$. So by induction, $(\pi \oplus \rho, \mu)_*(t) = \frac{1}{2}(0+0) = 0$. By Definition 1,

$$(\pi,\mu)_*(s) = (\pi,\mu)_*(t \frown a) = (\pi,\mu)_*(t)\pi(a|t) = 0\pi(a|t) = 0,$$

and similarly $(\rho,\mu)_*(s) = 0$ and $(\pi \oplus \rho,\mu)_*(s) = 0$. Thus $(\pi \oplus \rho,\mu)_*(s) = \frac{1}{2}((\pi,\mu)_*(s) + (\rho,\mu)_*(s))$ as desired.

Subcase 3.2: $\pi_*(t) + \rho_*(t) > 0$. Then

$$\begin{aligned}
(\pi \oplus \rho,\mu)_*(s) &= (\pi \oplus \rho,\mu)_*(t \frown a) \\
&= (\pi \oplus \rho,\mu)_*(t)(\pi \oplus \rho)(a|t) && \text{(Definition 1)} \\
&= (\pi \oplus \rho,\mu)_*(t)\frac{\pi_*(t \frown a) + \rho_*(t \frown a)}{\pi_*(t) + \rho_*(t)} && \text{(Definition 2)} \\
&= (\pi \oplus \rho,\mu)_*(t)\frac{\pi_*(t)\pi(a|t) + \rho_*(t)\rho(a|t)}{\pi_*(t) + \rho_*(t)} && \text{(Definition 1)} \\
&= \frac{(\pi,\mu)_*(t) + (\rho,\mu)_*(t)}{2}\frac{\pi_*(t)\pi(a|t) + \rho_*(t)\rho(a|t)}{\pi_*(t) + \rho_*(t)} && \text{(Induction)} \\
&= \frac{(\pi_*(t) + \rho_*(t))\mu_*(t)}{2}\frac{\pi_*(t)\pi(a|t) + \rho_*(t)\rho(a|t)}{\pi_*(t) + \rho_*(t)} && \text{(Lemma 1)} \\
&= (\pi_*(t)\mu_*(t)\pi(a|t) + \rho_*(t)\mu_*(t)\rho(a|t))/2 && \text{(Algebra)} \\
&= ((\pi,\mu)_*(t)\pi(a|t) + (\rho,\mu)_*(t)\rho(a|t))/2 && \text{(Lemma 1)} \\
&= ((\pi,\mu)_*(t \frown a) + (\rho,\mu)_*(t \frown a))/2, && \text{(Definition 1)}
\end{aligned}$$

as desired.

4 Duality and Janus agents: Strengthening and formalizing a result of Alexander and Hutter

Definition 3. *(Duality) Define $\overline{s}$, $\overline{\pi}$, $\overline{\mu}$ as in Reward-Punishment Symmetric Universal Intelligence.*

The following class of agents are named after Janus, the Roman god of duality, who features two faces, one facing forward, one facing backward.

Definition 4. *(Janus agents) By a* Janus agent*, we mean an agent π such that $\overline{\pi} = \pi$.*

Theorem 2. *Suppose Υ is a weighted performance averager. If $\Upsilon(\pi) = 0$ for every Janus agent π, then $\Upsilon(\overline{\pi}) = -\Upsilon(\pi)$ for every agent π.*

Proof. ...

5 Detectability of sets of agents

Definition 5. *A set Π of agents is* detectable *if there exists an environment μ such that for every agent π:*

1. *V_μ^π exists.*
2. *$V_\mu^\pi > 0$ if $\pi \in \Pi$.*
3. *$V_\mu^\pi \leq 0$ if $\pi \notin \Pi$.*

Definition 6. *A set Π is* closed under $\oplus$ *if the following condition holds: whenever $\pi \in \Pi$ and $\rho \in \Pi$, then $\pi \oplus \rho \in \Pi$.*

Theorem 3. *Let Π be any set of agents. If Π is detectable, then Π is closed under $\oplus$, and so is its complement Π^c.*

Proof. Assume Π is detectable. Let μ be as in Definition 5. To see Π is closed under $\oplus$, let $\pi, \rho \in \Pi$. Then $V_\mu^\pi > 0$ and $V_\mu^\rho > 0$, thus $V_\mu^{\pi\oplus\rho} > 0$ since $V_\mu^{\pi\oplus\rho} = \frac{1}{2}(V_\mu^\pi + V_\mu^\rho)$ by Theorem 1. So by choice of μ, $\pi \oplus \rho \in \Pi$. A similar argument shows that Π^c is closed under $\oplus$.

Definition 7. *An agent π is* Markov *if $\pi(a|s)$ only depends on the most recent observation in s.*

We will not use the following Lemma, but we state it in order to make it clearer what exactly a Markov agent is.

Lemma 4. *By a* policy *we mean a function f which takes a single observation $o \in \mathcal{O}$ as input, and outputs a probability distribution $a \mapsto f(a|o)$ on $\mathcal{A}$. For each policy f, let $\hat{f}$, the* agent defined by *f, be the agent defined as follows: for every s with most recent observation o, $\hat{f}(a|s) = f(a|o)$. Then: The set of Markov agents is exactly $\{\hat{f} : f \text{ is a policy}\}$.*

Proof. Straightforward.

Theorem 4. *The set of Markov agents is not detectable.*

Proof. Let Π be the set of Markov agents. By Theorem 3, it suffices to show Π is not closed under $\oplus$. Since $|\mathcal{A}| > 1$, we may choose two distinct $a_1, a_2 \in \mathcal{A}$. Define agents π, ρ by

$$\pi(a|s) = \begin{cases} .9 & \text{if } a = a_1, \\ .1 & \text{if } a = a_2, \\ 0 & \text{otherwise;} \end{cases}$$

$$\rho(a|s) = \begin{cases} .1 & \text{if } a = a_1, \\ .9 & \text{if } a = a_2, \\ 0 & \text{otherwise.} \end{cases}$$

Let p be any percept, let $s_1 = \langle p \rangle$, and let $s_2 = \langle p, a_1, p \rangle$. Then

$$(\pi \oplus \rho)(a_1|s_1) = \frac{\pi_*(s_1 \frown a_1) + \rho_*(s_1 \frown a_1)}{\pi_*(s_1) + \rho_*(s_1)} = \frac{0.9 + 0.1}{1 + 1} = \frac{1}{2}$$

and

$$(\pi \oplus \rho)(a_1|s_2) = \frac{\pi_*(s_2 \frown a_1) + \rho_*(s_2 \frown a_1)}{\pi_*(s_2) + \rho_*(s_2)} = \frac{0.9 \cdot 0.9 + 0.1 \cdot 0.1}{0.9 + 0.1} = \frac{82}{100} \neq \frac{1}{2}.$$

So $(\pi \oplus \rho)(a|s)$ depends on more than just the most recent observation in s, so $\pi \oplus \rho$ is not Markov.

Theorem 4 seems to cast doubt on Silver et al's informal conjecture [SSPS21] that the creation of strong enough RL agents is a direct pathway to Artificial General Intelligence (see [Ale21] for some other thoughts of ours on this conjecture). It seems reasonable to expect that an AGI should easily be capable of performing the task commanded by the command: "Please act in a Markov way." Theorem 4 suggests that despite this task's apparent simplicity (in contrast with playing chess or Go, for example), there is no way to *express* the task (or our desire for the AGI to perform it) using only the incentive structure of RL.

References

[AH21] Samuel Allen Alexander and Marcus Hutter. Reward-punishment symmetric universal intelligence. In *AGI*, 2021.

[Ale21] Samuel Allen Alexander. Can reinforcement learning learn itself? A reply to 'Reward is enough'. In *CIFMA*, 2021.

[SSPS21] David Silver, Satinder Singh, Doina Precup, and Richard Sutton. Reward is enough. *Artificial Intelligence*, 2021.