Reward-Punishment Symmetric Universal Intelligence

Samuel Alexander, SEC, samuelallenalexander@gmail.com

Marcus Hutter, DeepMind & ANU

Motivation

- Measuring intelligence is a key step toward AGI
- The Legg-Hutter universal intelligence measure is one approach, but it depends on choice of UTM
- Leike/Hutter (2015): “What are other desirable properties of a UTM?”
- We propose a symmetry constraint on UTMs

Review: Kolmogorov Complexity

Definition: For any prefix-free UTM U (and, implicitly, a suitable encoding of RL), the *Kolmogorov complexity* $K_U(\mu)$ of a computable RL environment μ is the length of the shortest U-computer program for μ.

Andrey Kolmogorov (1913-1987)

Review: Legg-Hutter Universal Intelligence

Definition: If U is a prefix-free UTM, the Legg-Hutter Universal Intelligence of RL agent π is

$$\Upsilon_U(\pi) = \sum_{\mu} 2^{-K_U(\mu)} V_{\mu}^{\pi}$$

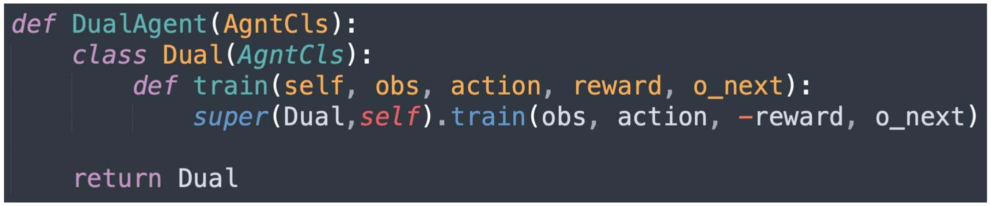
where V_{μ}^{π} is the total expected reward for π on μ.

William of Ockham (1285-1347)

Dual Agents and Environments

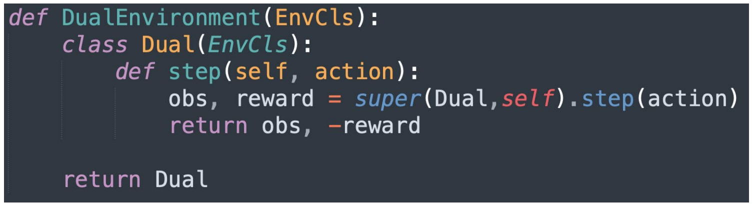
Definition: $\bar{\pi}$ is the agent which acts as π would act if π mistook rewards for punishments and punishments for rewards.

Definition: $\bar{\mu}$ is the environment which responds as μ would respond if μ mistook rewards for punishments and punishments for rewards.



```python
def DualAgent(AgntCls):
    class Dual(AgntCls):
        def train(self, obs, action, reward, o_next):
            super(Dual,self).train(obs, action, -reward, o_next)

    return Dual
```



```python
def DualEnvironment(EnvCls):
    class Dual(EnvCls):
        def step(self, action):
            obs, reward = super(Dual,self).step(action)
            return obs, -reward

    return Dual
```

Kolmogorov Complexity Symmetry

Definition: A prefix-free UTM U is *symmetric* if

$$K_U(\mu) = K_U(\bar{\mu})$$

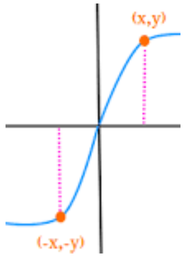
for every RL environment μ.

Theorem: Any prefix-free UTM can be transformed into a symmetric UTM under a mild technical assumption on how RL is encoded.

Universal Intelligence Symmetry

Theorem (symmetry about the origin): If U is symmetric then

$$\Upsilon_U(\bar{\pi}) = -\Upsilon_U(\pi).$$



Corollary: If U is symmetric and π ignores rewards then $\Upsilon_U(\pi) = 0$.

“But why should intelligence be symmetric?”

- Assume Υ measures intelligence as average performance.
- Say Υ is *weak symmetric* if: whenever $\Upsilon(\pi) \neq 0$ then $\Upsilon(\pi) \neq \Upsilon(\bar{\pi})$.

Weak symmetry is a reasonable/natural requirement: Say $\Upsilon(\pi)>0$. This should mean π is intelligent: π uses ingenuity to get positive rewards. By def., $\bar{\pi}$ uses that same ingenuity to obtain *punishments*. So it would be strange for $\bar{\pi}$ to get the *exact* same average rewards as π!

(We don’t state weak symmetry as absolute law, we merely opine it’s reasonable/natural)

Weak Symmetry implies Symmetry

Let π be any agent. Assume Υ is weak symmetric and measures intelligence as average performance.

Let ρ be an agent who, at the start of every environment, flips a coin and thereafter plays as π if HEADS, $\bar{\pi}$ if TAILS.

Since Υ measures avg. performance, $\Upsilon(\rho) = \frac{\Upsilon(\pi)+\Upsilon(\bar{\pi})}{2}$.

Define ρ' the same but swap HEADS and TAILS. ρ seems indistinguishable from ρ' so $\Upsilon(\rho) = \Upsilon(\rho')$. Swapping HEADS and TAILS is the same as swapping π and $\bar{\pi}$, thus ρ'= $\bar{\rho}$. Thus $\Upsilon(\rho) = \Upsilon(\bar{\rho})$. By weak symmetry, $\Upsilon(\rho) = 0$.

Thus $\Upsilon(\bar{\pi}) = -\Upsilon(\pi)$!

A few misc. notes

- Our existence proof of symmetric UTMs works by eliminating RL bias: RL is biased by the convention “+reward good, -reward bad”
- $|\Upsilon_U(\pi)|$ could be an alternate intelligence measure. Whereas $\Upsilon_U(\pi)$ measures performance, $|\Upsilon_U(\pi)|$ would measure ability to consistently extremize rewards (whether consistently positively or consistently negatively)
- Our stance is that $|\Upsilon_U(\pi)|$ and $\Upsilon_U(\pi)$ are equally valid intelligence measures which measure different aspects of intelligence
- It could be argued that the $|\Upsilon_U(\pi)|$ vs. $\Upsilon_U(\pi)$ debate goes all the way back to Plato’s “Lesser Hippias”

Summary

- Symmetric UTMs make symmetric Legg-Hutter intelligence measures.
- We argued symmetry is a reasonable intelligence-measure axiom.
- Symmetric UTMs can be built by eliminating a certain RL bias.
- This could narrow the space of UTMs, advancing AGI development.

Acknowledgments: José Hernández-Orallo, Shane Legg, Pedro Ortega, reviewers

Inherent Bias in RL

- RL is inherently biased because of its arbitrary convention that positive rewards are good and negative rewards are bad.

- Imagine a parallel universe where RL is formalized with positive rewards bad, negative rewards good. RL would work just as well.

- Our existence proof works by eliminating this bias: valid programs are required to state which RL convention they’re written for.

Whether to Use Absolute Values

- In a true-false IQ test, 0% is as hard to get as 100%.
- So then, if an agent consistently manages to get environments to punish it, shouldn’t we assume the agent is intelligent (but loves punishment)?
- Shouldn’t we use $|\Upsilon_U(\pi)|$ instead of $\Upsilon_U(\pi)$ to measure π’s intelligence?

Whether to use Absolute Values (cont’d)

Shouldn’t we use $|\Upsilon_U(\pi)|$ instead of $\Upsilon_U(\pi)$ to measure π’s intelligence?

- Our stance:
  - $\Upsilon_U(\pi)$ measures intelligence as average performance.
  - $|\Upsilon_U(\pi)|$ measures intelligence as ability to consistently extremize rewards (whether consistently positive or consistently negative).
  - We consider them equally valid intelligence measures. They measure different aspects of intelligence.

Abs Values History: Plato’s “Lesser Hippias”

- Whether to take absolute values is an ancient debate.
- In Plato’s “Lesser Hippias”, Socrates presents what initially seems like a compelling argument in favor of taking absolute values.

SOCRATES: “Which of the two then is a better runner? He who runs slowly voluntarily, or he who runs slowly involuntarily?” Etc. etc. etc...

Abs Values History: Socrates’ Evil Twist

- From what initially seems like a pro-abs-values argument, Socrates uses the same logic to defend the ludicrous position that it’s better to be intentionally evil than unintentionally evil.

(An interesting AGI safety question. Is an intentionally evil AGI better or worse than an unintentionally evil one?)

The dialogue ends with poor Hippias hopelessly confused. Better not to take sides on the abs-value question.