Symmetric Universal Intelligence: A Variation of Legg-Hutter Universal Intelligence

Samuel Allen Alexander

February 26, 2021

Abstract

Can an agent's intelligence level be negative? We extend the Legg-Hutter agent-environment framework to include environments capable of punishing the agent with negative rewards. The Legg-Hutter intelligence of an agent attempts to measure how much reward the agent extracts from the universe of all environments in an aggregate sense. Thus, when the framework is extended to allow punishments, it naturally follows that an agent should have negative intelligence if that agent generally extracts more punishments than rewards. Extending the framework in this way, we discover algebraic structural properties not otherwise present. We introduce a variation of Legg-Hutter universal intelligence, which we call symmetric universal intelligence, and use said structural properties to show that this symmetric universal intelligence measure satisfies certain intuitive algebraic desiderata.

1 Introduction

In their brilliant paper [8], Legg and Hutter write:

> "As our goal is to produce a definition of intelligence that is as broad and encompassing as possible, the space of environments used in our definition should be as large as possible."

So motivated, we investigated what would happen if we extended the universe of environments to include environments with rewards from $\mathbb{Q} \cap [-1, 1]$ instead of just from $\mathbb{Q} \cap [0, 1]$ as in Legg and Hutter's paper. In other words, we investigated what would happen if environments are not only allowed to reward agents but also to punish agents (a punishment being a negative reward).

We discovered that when negative rewards are allowed, this introduces a certain algebraic structure into the agent-environment framework. The main objection we anticipate to our extended framework is that it implies the negative intelligence of certain agents[1]. We would argue that this makes perfect sense when environments are capable of punishing agents: the intelligence of

[1]Thus, this paper falls under the broader program of advocating for intelligence measures

a reinforcement learning agent should measure the degree to which that agent can extract large rewards on average across many environments; an agent who instead extracts large punishments on average across many environments should therefore have a negative intelligence level.

The structure of the paper is as follows:

- In Section 2, we give preliminary definitions.
- In Section 3, we introduce what we call the dual of an agent and of an environment, and prove some algebraic theorems about these.
- In Section 4, we introduce symmetric universal intelligence, a variation of Legg-Hutter intelligence, and show it satisfies certain algebraic desiderata.
- In Section 5, we summarize and make concluding remarks, including remarks about how these ideas might be applied to certain other intelligence measures.

2 Preliminaries

In defining agent and environment below, we attempt to follow Legg and Hutter [8] as closely as possible, except that we permit environments to output rewards from $\mathbb{Q} \cap [-1, 1]$ rather than just $\mathbb{Q} \cap [0, 1]$ (and, accordingly, we modify which well-behaved environments to restrict our attention to).

We work in a background theory of computability given by a prefix-free universal Turing machine. Throughout the paper, we implicitly fix a finite set $\mathcal{A}$ of *actions*, a finite set $\mathcal{O}$ of *observations*, and a finite set $\mathcal{R} \subseteq \mathbb{Q} \cap [-1, 1]$ of *rewards* (so each reward is a rational number between -1 and 1 inclusive), with $|\mathcal{A}| > 1$, $|\mathcal{O}| > 0$, $|\mathcal{R}| > 0$. We assume that $\mathcal{R}$ satisfies the following property: whenever $\mathcal{R}$ contains any reward r, then $\mathcal{R}$ also contains $-r$. We assume $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{R}$ are mutually disjoint (i.e., no reward is an action, no reward is an observation, and no action is an observation). The acronym ORA stands for "Observation-Reward-Action". By $\langle\rangle$ we mean the empty sequence.

Definition 1. *(Agents, environments, etc.)*

1. *By an* ORA-sequence*, we mean a sequence s such that the following requirements hold for every natural number n:*
   - *If s has a $(3n)$th element, that element is an observation.*
   - *If s has a $(3n+1)$th element, that element is a reward.*
   - *If s has a $(3n+2)$th element, that element is an action.*

having different ranges than the nonnegative reals. In other papers we have advocated more extreme extensions of the range of intelligence measures [2]; by contrast, here we merely question the assumption that intelligence never be negative, leaving aside the question of whether intelligence should be real-valued.

Thus, informally speaking, an ORA-sequence is empty or looks like "observation, reward, action, observation, reward, action, ..."; if it terminates (i.e., if it is finite), it may terminate with either an observation, a reward, or an action.

2. *By an* ORA-prompt*, we mean a non-empty ORA-sequence which terminates with a reward.*

3. *By an* ORA-play*, we mean an ORA-sequence which is either empty or else terminates with an action.*

4. *By an* agent*, we mean a function π which assigns to every ORA-prompt p a probability measure $\pi(p)$ on $\mathcal{A}$. For each $a \in \mathcal{A}$, we write $\pi(a|p)$ for the probability assigned to a by $\pi(p)$. Intuitively, we think of $\pi(a|p)$ as the probability that the agent π will take action a in response to the ORA-prompt p.*

5. *By an* environment*, we mean a function μ which assigns to every ORA-play p a probability measure $\mu(p)$ on $\mathcal{O} \times \mathcal{R}$. For each $o \in \mathcal{O}$ and $r \in \mathcal{R}$, we write $\mu(o, r|p)$ for the probability assigned to (o, r) by $\mu(p)$. Intuitively, we think of $\mu(o, r|p)$ as the probability that the environment μ will issue observation o and reward r to the agent in response to the ORA-play p.*

6. *If π is an agent, μ is an environment, and $n \in \mathbb{N}$, we write $V^\pi_{\mu,n}$ for the expected value of the sum of the first n rewards which would occur in an ORA-sequence $(o_0, r_0, a_0, \ldots, o_n, r_n)$ randomly generated as follows:*

   (a) *$(o_0, r_0) \in \mathcal{O} \times \mathcal{R}$ is chosen randomly based on the probability measure $\mu(\langle\rangle)$.*

   (b) *$a_0 \in \mathcal{A}$ is chosen randomly based on the probability measure $\pi(o_0, r_0)$.*

   (c) *For each $i > 0$, $(o_i, r_i) \in \mathcal{O} \times \mathcal{R}$ is chosen randomly based on the probability measure $\mu(o_0, r_0, a_0, \ldots, o_{i-1}, r_{i-1}, a_{i-1})$.*

   (d) *For each $i > 0$, $a_i \in \mathcal{A}$ is chosen randomly based on the probability measure $\pi(o_0, r_0, a_0, \ldots, o_{i-1}, r_{i-1}, a_{i-1}, o_i, r_i)$.*

7. *If π is an agent and μ is an environment, let $V^\pi_\mu = \lim_{n\to\infty} V^\pi_{\mu,n}$. Intuitively, V^π_μ is the expected total reward which π would extract from μ.*

Note that it is possible for V^π_μ to be undefined. For example, if μ is an environment which always issues reward $(-1)^n$ in response to the agent's nth action, then V^π_μ is undefined for every agent π. This would not be the case if rewards were required to be ≥ 0, so this is one way in which allowing punishments complicates the resulting theory.

In order to define Legg-Hutter-style intelligence measures, it is, in any case, necessary to restrict attention to certain well-behaved environments. Legg and Hutter (in their Section 3.2) restrict attention to environments μ that never give total reward more than 1, from which it follows that $V^\pi_\mu \leq 1$. This implication is less clear when environments can punish the agent, so we take a different

approach: rather than limit the total reward the environment can give and use that to force V_μ^π to be bounded, we will cut the middleman and directly assume that V_μ^π is bounded.

Definition 2. *A environment μ is* well-behaved *if μ is computable and the following condition holds: for every agent π, V_μ^π exists and $-1 \leq V_\mu^\pi \leq 1$.*

3 Dual Agents and Dual Environments

In the Introduction, we promised that by allowing environments to punish agents, we would reveal algebraic structure not otherwise present. The key to this additional structure is the following definition.

Definition 3. *(Dual Agents and Dual Environments)*

1. *Suppose s is an ORA-sequence. Let $-s$ be the ORA-sequence which is equal to s in every way except that, whenever the $(3n+1)$th element of s is a reward r, then the $(3n+1)$th element of $-s$ is $-r$.*

2. *Suppose π is an agent. We define a new agent $-\pi$, the* dual *of π, as follows: for each ORA-prompt p, for each action $a \in \mathcal{A}$,*

$$(-\pi)(a|p) = \pi(a|-p).$$

3. *Suppose μ is an environment. We define a new environment $-\mu$, the* dual *of μ, as follows: for each ORA-play p, for each observation $o \in \mathcal{O}$ and reward $r \in \mathcal{R}$,*

$$(-\mu)(o, r|p) = \mu(o, -r|-p).$$

Lemma 4. *(Double Negation)*

1. *For each ORA-sequence s, $--s = s$.*

2. *For each agent π, $--\pi = \pi$.*

3. *For each environment μ, $--\mu = \mu$.*

Proof. Follows from the fact that for every real number r, $--r = r$. □

In the proof of the following theorem, we write $\frown$ for concatenation.

Theorem 5. *Suppose μ is an environment and π is an agent. Then*

$$V_{-\mu}^{-\pi} = -V_\mu^\pi$$

(and the left-hand side is defined if and only if the right-hand side is defined).

Proof. By Definition 1 part 7, it suffices to show that for each $n \in \mathbb{N}$, $V^{-\pi}_{-\mu,n} = -V^{\pi}_{\mu,n}$. For that, it suffices to show that for every finite ORA-sequence p, if p is empty or p terminates with a reward or an action, then the probability X of generating p using π and μ (as in Definition 1 part 6) equals the probability X' of generating $-p$ using $-\pi$ and $-\mu$. We will show this by induction on the length of p.

Case 1: p is empty. Then $X = X' = 1$.

Case 2: p terminates with a reward. Then $p = q \frown o \frown r$ for some ORA-sequence q which terminates with an action. Let Y (resp. Y') be the probability of generating q (resp. $-q$) using π and μ (resp. $-\pi$ and $-\mu$). We reason:

$$\begin{aligned} X &= \mu(o, r|q)Y && \text{(Definition of } X\text{)} \\ &= \mu(o, - - r|- - q)Y && \text{(Lemma 4)} \\ &= (-\mu)(o, -r|-q)Y && \text{(Definition of } -\mu\text{)} \\ &= (-\mu)(o, -r|-q)Y' && \text{(By induction, } Y = Y'\text{)} \\ &= X'. && \text{(Definition of } X'\text{)} \end{aligned}$$

Case 3: p terminates with an action. Then $p = q \frown a$ for some ORA-sequence q which terminates with a reward. Let Y (resp. Y') be the probability of generating q (resp. $-q$) using π and μ (resp. $-\pi$ and $-\mu$). We reason:

$$\begin{aligned} X &= \pi(a|q)Y && \text{(Definition of } X\text{)} \\ &= \pi(a|- - q)Y && \text{(Lemma 4)} \\ &= (-\pi)(a|-q)Y && \text{(Definition of } -\pi\text{)} \\ &= (-\pi)(a|-q)Y' && \text{(By induction, } Y = Y'\text{)} \\ &= X'. && \text{(Definition of } X'\text{)} \end{aligned}$$

□

Corollary 6. *For every agent π and environment μ,*

$$V^{\pi}_{-\mu} = -V^{-\pi}_{\mu}$$

(and the left-hand side is defined if and only if the right-hand side is defined).

Proof. If neither side is defined, then there is nothing to prove. Assume the left-hand side is defined. Then

$$\begin{aligned} V^{\pi}_{-\mu} &= V^{--\pi}_{-\mu} && \text{(Lemma 4)} \\ &= -V^{-\pi}_{\mu}, && \text{(Theorem 5)} \end{aligned}$$

as desired. A similar argument holds if we assume the right-hand side is defined. □

Corollary 7. *For every environment μ, μ is well-behaved if and only if $-\mu$ is well-behaved.*

Proof. We prove the $\Rightarrow$ direction, the other is similar. Since μ is well-behaved, μ is computable, so clearly $-\mu$ is computable. Let π be any agent. Since μ is well-behaved, $V_\mu^{-\pi}$ is defined and $-1 \leq V_\mu^{-\pi} \leq 1$. By Corollary 6, $V_{-\mu}^{\pi} = -V_\mu^{-\pi}$ is defined, implying $-1 \leq V_{-\mu}^{\pi} \leq 1$. By arbitrariness of π, this shows $-\mu$ is well-behaved. □

4 Symmetric Intelligence

By definition, agent $-\pi$ acts exactly as agent π would act if π were confused into thinking that punishments were rewards and rewards were punishments. Whatever ingenuity π applies to maximize rewards, $-\pi$ applies that exact same ingenuity to maximize punishments. Thus, if Γ is an intelligence function (intended to measure how well an agent extracts rewards in some aggregate sense across the whole infinite space of all well-behaved environments), then Γ ought to satisfy the equation:

$$\Gamma(-\pi) = -\Gamma(\pi).$$

Not every intelligence function satisfies the above equation, but the above equation should be considered desirable when we invent ways of measuring intelligence: all else being equal, an intelligence function which satisfies the above equation should be considered better than an intelligence function which does not[2]. In this section, we will argue that the universal intelligence measure proposed by Legg and Hutter can be varied in a natural way which will make it satisfy the above equation.

Definition 8. *(Complexity) Suppose μ is an environment.*

1. *By $K(\mu)$, we mean the Kolmogorov complexity of μ.*
2. *By $S(\mu)$, the* symmetric complexity *of μ, we mean*

$$S(\mu) = \max\{K(\mu), K(-\mu)\}.$$

Lemma 9. *For any environment μ, $S(\mu) = S(-\mu)$.*

Proof. Trivial. □

Definition 10. *Let W be the set of all well-behaved environments. Let $-W = \{-\mu : \mu \in W\}$.*

[2]This depends strongly on our assumption that when we measure intelligence we are measuring aggregate reward-extraction. This might clash with everyday ideas of intelligence: if a mischievous child answers every single question wrong on a true-false IQ test, we would say that child is clearly very intelligent, and is using said intelligence to deliberately fail the test. To quote Socrates, "Don't you think the ignorant person would often involuntarily tell the truth when he wished to say falsehoods, if it so happened, because he didn't know; whereas you, the wise person, if you should wish to lie, would always consistently lie?" [10] Likewise if an agent consistently manages to extract the worst possible rewards from many environments, we might think the agent is brilliant but masochistic. Nevertheless, we would not want to put said agent in charge of navigating environments for us (at least not without some reverse psychology).

Lemma 11. $W = -W$.

Proof. By Corollary 7. □

Definition 12. *(Universal Intelligence) Suppose π is an agent.*

1. *The* Legg-Hutter universal intelligence $\Gamma_{LH}(\pi)$ *is*

$$\Gamma_{LH}(\pi) = \sum_{\mu \in W} 2^{-K(\mu)} V_\mu^\pi.$$

2. *The* symmetric universal intelligence $\Gamma_S(\pi)$ *is*

$$\Gamma_S(\pi) = \sum_{\mu \in W} 2^{-S(\mu)} V_\mu^\pi.$$

The sum defining Γ_{LH} is absolutely convergent because the summands in absolute value are dominated by the summands defining Chaitin's constant (this is why in Section 2 we assume a background model of computation given by a *prefix-free* universal Turing machine). Thus the sum defining Γ_S is also absolutely convergent since $2^{-S(\mu)} \leq 2^{-K(\mu)}$. Thus, by a well-known theorem from elementary calculus, the sums do not depend on which order the $\mu \in W$ are enumerated.

The following theorem shows that Γ_S satisfies the above desideratum.

Theorem 13. *For every agent π,*

$$\Gamma_S(-\pi) = -\Gamma_S(\pi).$$

Proof. Let π be an agent. Then:

$$\begin{aligned}
\Gamma_S(-\pi) &= \sum_{\mu \in W} 2^{-S(\mu)} V_\mu^{-\pi} && \text{(Definition 12)} \\
&= -\sum_{\mu \in W} 2^{-S(\mu)} V_{-\mu}^{\pi} && \text{(Corollary 6)} \\
&= -\sum_{\mu \in W} 2^{-S(-\mu)} V_{-\mu}^{\pi} && \text{(Lemma 9)} \\
&= -\sum_{\mu \in -W} 2^{-S(\mu)} V_{\mu}^{\pi} && \text{(Change of variables)} \\
&= -\sum_{\mu \in W} 2^{-S(\mu)} V_{\mu}^{\pi} && \text{(Lemma 11)} \\
&= -\Gamma_S(\pi). && \text{(Definition 12)}
\end{aligned}$$

□

The above desideratum, that $\Gamma(-\pi) = -\Gamma(\pi)$, applies to numerical intelligence measures. If one is merely interested in binary intelligence comparators (such as those in [1]), the desideratum can be weakened into a non-numerical comparator form:

If π is more intelligent than ρ, then $-\pi$ should be less intelligent than $-\rho$.

The following corollary shows that Γ_S satisfies this desideratum too.

Corollary 14. *Let π, ρ be agents. If $\Gamma_S(\pi) > \Gamma_S(\rho)$ then $\Gamma_S(-\pi) < \Gamma_S(-\rho)$.*

Proof. By Theorem 13 and basic algebra. □

The following corollary shows that Γ_S satisfies another obvious desideratum.

Corollary 15. *Suppose π is an agent which ignores rewards (by which we mean that $\pi(p)$ does not depend on the rewards in p). Then $\Gamma_S(\pi) = 0$.*

Proof. The hypothesis clearly implies $\pi = -\pi$. Thus by Theorem 13, $\Gamma_S(\pi) = -\Gamma_S(\pi)$, which forces $\Gamma_S(\pi) = 0$. □

Corollary 15 suggests symmetric complexity may be an improvement over Kolmogorov complexity in Definition 12. Let $a \in \mathcal{A}$ be some default action and consider the agent π_a who always takes action a. For any particular environment μ, where π_a earns some total reward r by blind luck, that total reward ought to be cancelled out by $-\mu$, where the exact same blind luck becomes blind misfortune and π_a earns total reward $-r$ (Corollary 6). But if $K(\mu) \neq K(-\mu)$, then the different weights $2^{-K(\mu)} \neq 2^{-K(-\mu)}$ would prevent these two outcomes from cancelling each other. Similarly, if an environment μ ignores agent actions, Corollary 6 implies the contributions of μ and $-\mu$ cancel each other out in $\Gamma_S(\pi)$, which is desirable because why should an agent's intelligence depend on its interaction with an environment that ignores it?

Of course, Kolmogorov complexity and hence Legg-Hutter intelligence depends implicitly on the background model of computation [9]. One could contrive a background model of computation where $K = S$, in which case the distinction in Definition 12 would disappear.

5 Conclusion

By extending our attention to environments which can punish agents, we became aware of algebraic structure in the agent-environment framework (Section 3). This structure allowed us to show that by varying the weights used in the Legg-Hutter universal intelligence construction, we can construct a similar intelligence measure which satisfies key desiderata (Theorem 13, Corollaries 14 and 15).

Similar modifications could be made to other intelligence measures based on the original Legg-Hutter idea, such as the anytime intelligence measure of Hernández-Orallo and Dowe [5] and perhaps the realtime intelligence measure of Gavane [4], although the latter may involve greater care since negating numbers takes a nonzero amount of time. The precise strategy we employ in this paper is not relevant to Hibbardian intelligence measures of agents who compete in adversarial sequence prediction games [6] [3]. However, the higher-level idea still applies: a predictor who always dumbly makes the same prediction should have

intelligence 0, and yet, in some sense such a predictor still outperforms predictors who intentionally try to lose, and thus one could argue that intentionally-losing predictors really ought to have negative Hibbardian intelligence.

Beyond the technical details, at a higher level, a key takeaway of this paper is that if intelligence is supposed to measure how much reward an agent extracts in aggregate across many environments, including environments capable of punishing the agent, then such intelligence measures ought to be permitted to output negative values for certain agents. It is fun to imagine the trouble that maximally anti-intelligent agents, such as the negative version of the universally intelligent agent AIXI [7], would get themselves into.

References

[1] Samuel Allen Alexander. Intelligence via ultrafilters: structural properties of some intelligence comparators of deterministic Legg-Hutter agents. *Journal of Artificial General Intelligence*, 10(1):24–45, 2019.

[2] Samuel Allen Alexander. The Archimedean trap: Why traditional reinforcement learning will probably not yield AGI. *Journal of Artificial General Intelligence*, 11(1):70–85, 2020.

[3] Samuel Allen Alexander and Bill Hibbard. Measuring intelligence and growth rate: Variations on Hibbard's intelligence measure. *Journal of Artificial General Intelligence*, 12(1):1–25, 2021.

[4] Vaibhav Gavane. A measure of real-time intelligence. *Journal of Artificial General Intelligence*, 4(1):31–48, 2013.

[5] José Hernández-Orallo and David L Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539, 2010.

[6] Bill Hibbard. Measuring agent intelligence via hierarchies of environments. In *ICAGI*, pages 303–308, 2011.

[7] Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer, 2004.

[8] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4):391–444, 2007.

[9] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. In *Conference on Learning Theory*, pages 1244–1259. PMLR, 2015.

[10] Plato. Lesser Hippias. In John M Cooper, Douglas S Hutchinson, et al., editors, *Plato: complete works*. Hackett Publishing, 1997.