THE GRAIN OF TRUTH PROBLEM

07/31/2024

Cole Wyeth,
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Work joint with Marcus Hutter

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

Relying on ideas from many others...

Multi-player Games

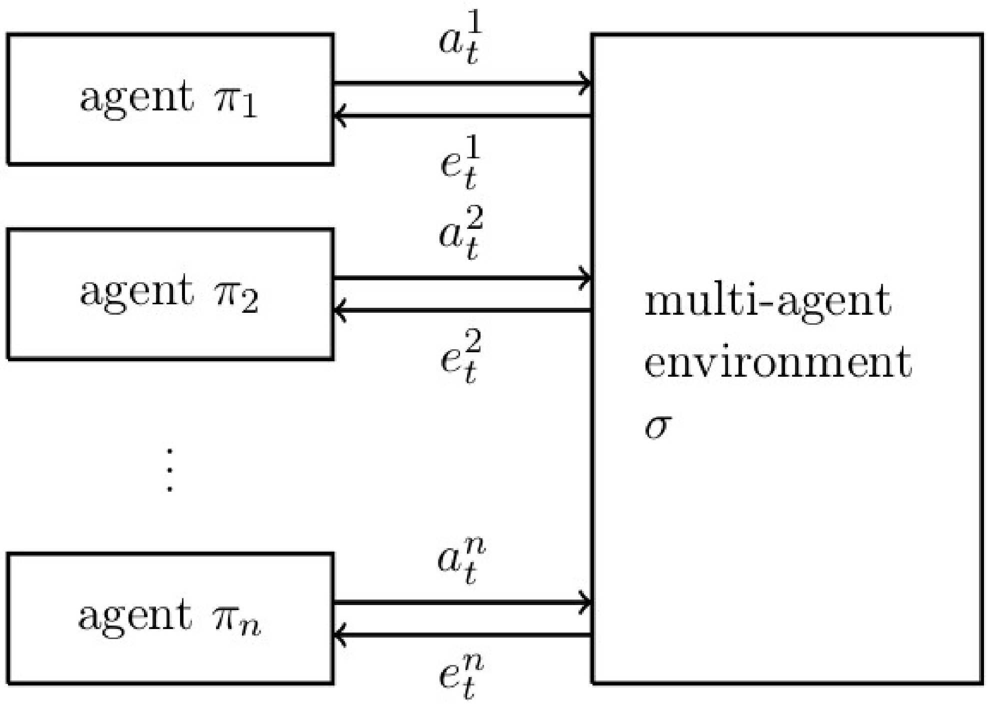

Figure 7.2: Agents $\pi_1, \ldots, \pi_n$ interacting in a multi-agent environment.

$$h_{1:t} = a_1 e_1 a_2 e_2 \ldots a_t e_t$$

$$h \sim \sigma^\pi$$

$$h^i_{1:t} := a^i_1 e^i_1 a^i_2 e^i_2 \ldots a^i_t e^i_t$$

$$h^i \sim \sigma^\pi_i$$

Multi-player Games

$$\sigma^\pi(\epsilon) = 1$$

$$\sigma^\pi(h_{<t}a_t) = \sigma^\pi(h_{<t}) \prod_{i=1}^{n} \pi_i(a_t^i | h_{<t}^i)$$

$$\sigma^\pi(h_{<t}a_te_t) = \sigma^\pi(h_{<t}a_t)\sigma(e_t | h_{<t}a_t)$$

$$\sigma_i^\pi(h_{<t}^i) = \sum_{h_{<t}^j, j \neq i} \sigma^\pi(h_{<t})$$

The Grain of Truth Problem

In a multi-player game, we would like all players to have a prior over the strategies they will face, and not be "infinitely surprised" by what actually happens.

We want to find

- a class of games $\mathcal{G}$
- a class of strategies $\mathcal{P}$
- priors ξ_i for each player i

So that each players optimal strategy $\pi^*_{\xi_i} \in \mathcal{P}$,

and $\xi_i \gg \sigma_i^\pi (\forall \sigma \in \mathcal{G}, \forall \pi \in \mathcal{P})$

Because we read "An Introduction to Universal Artificial Intelligence" [1] we want $\mathcal{G}$,$\mathcal{P}$ to at least include all computable games/strategies.

The Grain of Truth Problem

Note that in particular $\xi_i \gg \sigma_i^{\pi^*}$ where $\pi^* := (\pi^*_{\xi_1}, \pi^*_{\xi_2}, \ldots, \pi^*_{\xi_n})$

Learning occurs when $\mathcal{G}$, $\mathcal{P}$ are not singleton sets

A player's prior distribution ξ_i is over his subjective history, so that he is uncertain of his subjective environment; if the game is known, uncertainty is over the other players' actions/strategies instead

Simple (Non)Examples

- Any Nash equilibrium
- Prisoner's dilemma with generalized grim trigger strategies
- NOT the sets of computable (l.s.c., estimable) games and strategies
  - Diagonalization arguments

The Problem of Mutual Recursion

Imagine that we are about to play a simple game:

- I will secretly choose one of two cups to poison
- You will choose which cup to drink from

What you should do depends on which cup you think that I think that you think … that I poisoned.

But obviously playing the Nash equilibrium is reasonable

The Problem of Mutual Recursion

Reflective oracles solve this problem by “letting the recursion terminate in a fixed point.” Let $\lambda_T^O(\alpha|x) := P[T^O(x) = \alpha]$

Definition 7 (reflective oracle) An oracle $O : \mathcal{T} \times \mathcal{A}^* \times (\mathbb{Q} \cap [0,1]) \times \mathcal{A} \to [0,1]$ is called reflective iff for each pTM T and string $x \in \Sigma^*$, $\exists \{q_\alpha\}_{\alpha \in \mathcal{A}}$ satisfying the following properties:

$$\sum_{\alpha \in \mathcal{A}} q_\alpha = 1 \qquad (2)$$

And for all $\alpha \in \mathcal{A}$ and $p \in \mathbb{Q}$,

$$\lambda_T^O(\alpha|x) \le q_\alpha \le 1 - \sum_{\beta \neq \alpha} \lambda_T^O(\beta|x)$$

$$O_\alpha(T, x, p) = 1 \quad \text{for} \quad p < q_\alpha$$

$$O_\alpha(T, x, p) = 0 \quad \text{for} \quad p > q_\alpha$$

The Problem of Mutual Recursion

Why care about reflective oracles?

- Not only do reflective oracles exist, there are limit computable examples
- This means that all strategies we will discuss have “anytime algorithms”
- We extend existence and computability results to nonbinary and typed oracles

Convergence to Nash for Bayesian Players

Assuming σ is an infinitely repeated stage game, an old result of Kalai and Lehrer [4] shows that optimal players with priors satisfying the grain of truth property converge to a ε-Nash equilibrium:

THEOREM 2: *Let f and $f^1, f^2, \ldots, f^n$ be strategy vectors representing respectively the one actually played and the beliefs of the players. Suppose that for every player i:*

(i) *f_i is a best response to f^i_{-i}; and*

(ii) *f is absolutely continuous with respect to f^i.*

Then for every $\varepsilon > 0$ and for almost all (with respect to μ_f) play paths z there is a time $T = T(z, \varepsilon)$ such that for every $t \geqslant T$ there exists an ε-equilibrium $\bar{f}$ of the repeated game satisfying $f_{z(t)}$ plays ε-like $\bar{f}$.

Convergence to Nash for Bayesian Players

In our notation:

- The known game σ must be an infinitely repeated stage game
- The players must have independent priors over each opponent's strategy $\pi^i = (\pi^i_1, \ldots, \pi_i, \ldots, \pi^i_n)$
- The subjective environment prior is $\xi_i = \sigma_i^{\pi^i}$
- Then the grain of truth condition is sufficient for convergence to ε -Nash equilibrium

Convergence to Nash for Bayesian Players

Let $\lambda_T^O(\alpha|x)$ be the probability that probabilistic Turing machine T with access to O returns symbol α on input x.

We can "complete" λ_T^O to a measure $\pi_T := \bar{\lambda}_T^O$ by performing a binary search with O.

This constructs a computably enumerable policy class:

$$\mathcal{P}_{\text{refl}}^O := \{\pi_T\}_{T \in \mathcal{T}}$$

For now we will assume a known and infinitely repeated stage game, so that

$$\mathcal{G} := \{\sigma\}$$

Convergence to Nash for Bayesian Players

We construct a prior over opponent strategies with a recursive trick (become the prior you have been waiting for):

Algorithm 1 pTM Q

Input: History $æ_{<t}$
Require: Random sequence ω
Output: $a_t \sim \lambda_Q^O(a_t|æ_{<t})$

1: Obtain $\langle Q \rangle$
2: Let $\phi_\alpha(æ_{<t}, \cdot)$ approximate $\sum_{\pi \in \mathcal{P}_{\text{refl}}^O} w_\pi \frac{\pi(a_{<t}||e_{<t})}{\pi_Q(a_{<t}||e_{<t})} \pi(\alpha|æ_{<t})$ from below, where $\pi_Q \equiv \bar{\lambda}_Q^O$
3: Run sample(ϕ_α, $æ_{<t}$) with access to ω (Algorithm 7).

This is a “mixed strategy” in the sense of Kuhn’s theorem

In a known game, this immediately yields

$\xi_i := \sigma_i^{\pi^i}$ where $\pi^i := (\pi_Q, \ldots, \pi^*_{\xi_i}, \ldots, \pi_Q)$ so $\xi_i \gg \sigma_i^{\pi^*}$

Convergence to Nash for Bayesian Players

A reflective-oracle computable prior gives a reflective-oracle computable optimal value function:

$$V_\nu^*(h_{<t}a_t) \;=\; \frac{1}{\Gamma_t} \lim_{T\to\infty} \sum_{e_t} \max_{a_{t+1}} \sum_{e_{t+1}} \ldots \max_{a_T} \sum_{e_T} \sum_{i=t}^{T} \gamma_i r_i \prod_{j=t}^{T} \nu(e_j | h_{<t}æ_{<j}a_j)$$

Convergence to Nash for Bayesian Players

Further cleverness to break ties gives a reflective-oracle computable optimal strategy:

$$\lambda^{O}_{T_{\alpha\beta}}(\alpha|h_{<t}) \;=\; \frac{1}{2}[V^*_\nu(h_{<t}\alpha) - V^*_\nu(h_{<t}\beta) + 1] \;\in\; [0,1]$$

$$\lambda^{O}_{T_{\alpha\beta}}(\beta|h_{<t}) \;=\; 1 - \lambda^{O}_{T_{\alpha\beta}}(\alpha|h_{<t}) \;=\; \frac{1}{2}[V^*_\nu(h_{<t}\beta) - V^*_\nu(h_{<t}\alpha) + 1] \;\in\; [0,1]$$

$$\pi(a|h_{<t}) \;=\; \begin{cases} 1 \text{ if } a = \alpha \text{ and } O(T_{\alpha\beta}, h_{<t}, 1/2) \to 1, \\ 1 \text{ if } a = \beta \text{ and } O(T_{\alpha\beta}, h_{<t}, 1/2) \to 0, \\ 0 \text{ otherwise.} \end{cases}$$

Convergence to Nash for Bayesian Players

The result of Kalai and Lehrer shows that optimal play with these priors converges to a ε-Nash equilibrium

THEOREM 2: *Let f and $f^1, f^2, \ldots, f^n$ be strategy vectors representing respectively the one actually played and the beliefs of the players. Suppose that for every player i:*
(i) *f_i is a best response to f^i_{-i}; and*
(ii) *f is absolutely continuous with respect to f^i.*
Then for every $\varepsilon > 0$ and for almost all (with respect to μ_f) play paths z there is a time $T = T(z, \varepsilon)$ such that for every $t \geqslant T$ there exists an ε-equilibrium $\bar{f}$ of the repeated game satisfying $f_{z(t)}$ plays ε-like $\bar{f}$.

This means that after being poisoned enough times, Bayesians will eventually choose a (uniformly) random cup!

(Assuming the discount factor is low enough)

UNFORTUNATELY...

“Nobody cares about infinitely repeated stage games.”

MARCUS HUTTER

Unknown and General Games

Simple recipe:

- Introduce a “type system” to the reflective oracle (a distinct probability simplex for each players’ action space and for the percept space)
- Construct an explicit environment mixture (analogous to π_Q)
- Apply Thompson sampling instead of the optimal strategy

Algorithm 2 Thompson sampling strategy π_{TS}

Input: Percept stream $e_{1:\infty}$

Output: $a_{1:\infty} \sim \pi_{TS}(\cdot || e_{1:\infty})$

```
while true do
    sample ρ ~ w(·|æ_{<t})
    follow π*_ρ for H_t(ε_t) steps
```

An Application to Self-Predictive A.G.I.

Self-Predictive Universal AI

Elliot Catt, Jordi Grau Moya, Marcus Hutter, Matthew Aitchison
Tim Genewein, Gregoire Deletang, Kevin Li Wenliang, Joel Veness
Google DeepMind
ecatt@google.com

Abstract

Reinforcement Learning (RL) algorithms typically utilize learning and/or planning techniques to derive effective policies. Integrating both approaches has proven to be highly successful in addressing complex sequential decision-making challenges, as evidenced by algorithms such as AlphaZero and MuZero, which consolidate the planning process into a parametric search-policy. AIXI, the universal Bayes-optimal agent, leverages planning through comprehensive search as its primary means to find an optimal policy. Here we define an alternative universal Bayesian agent, which we call Self-AIXI, that on the contrary to AIXI, maximally exploits learning to obtain good policies. It does so by *self-predicting* its own stream of action data, which is generated, similarly to other TD(0) agents, by taking an action maximization step over the current on-policy (universal mixture-policy) Q-value estimates. We prove that Self-AIXI converges to AIXI, and inherits a series of properties like maximal Legg-Hutter intelligence and the self-optimizing property.

1 Introduction

Reinforcement Learning (RL) [1] algorithms exploit learning, planning [1], or their combination, to obtain good policies from experience. Pure learning consists of using real experience for improving a policy via a (parametric) model, possibly representing an explicit policy-model and/or the Q-values [2–6]. In a sense, learning stores the computational effort of policy-improvement into the parameters, which makes it a computationally efficient approach when needing to reuse the policy later on. In contrast, pure planning finds good policies via simulated experience using an environment model and a randomized (or exhaustive) search policy [1, 7, 8]. In the case of unknown or stochastic environments, one must re-plan after receiving a new observation, thus wasting all computational-effort from the previous step. This makes pure planning a wasteful approach. Using both, planning and learning, is a good way to improve performance and efficiency as demonstrated by modern high-performant RL algorithms such as MuZero [9–12]. These algorithms distill the planning effort back into the parametric search-policy by training it to predict the good actions obtained from planning. In a way, these agents are *self-predicting* their own policy-improvements. Although empirically successful and widely used, this distillation [13] or self-prediction [2] process is motivated in a purely heuristic way without much theoretical understanding on its optimality condition.

The AIXI agent [14, 15] is a theoretical universal Bayes-optimal agent obtained through pure planning without relying on distilling the search effort as described above. AIXI learns an environment model via a Solomonoff predictor [16, 17] and uses it for exhaustive (computationally intractable) planning. Thus, although it uses learning for the environment model, we say AIXI adopts a pure planning approach in the context of policy generation. Two desirable properties of Solomonoff prediction are universality—obtained by considering a huge hypothesis class containing all computable

[1] We use the terms planning and search interchangeably.

[2] Policy distillation usually refers to the process of amortizing one or several policies into another policy model. We view self-prediction as a type of distillation where a search-policy is consolidated into another model.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

What if you are uncertain about both the environment and your own policy?

Use environment mixture $\xi \geq \mathcal{M}$

And now a policy mixture $\zeta \geq \mathcal{P}$

$$\pi_S(h_{<t}) := \arg\max_{a_t} Q_\xi^\zeta(h_{<t}, a_t)$$

Hope that $\pi_S \to \pi_\xi^*$

A helpful condition would be $\pi_S \in \mathcal{P}$

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

An Application to Self-Predictive A.G.I.

Choose the classes of reflective-oracle computable policies and environments

$$\pi_S(h_{<t}) \;\in\; \operatorname{argmax}_{a_t \in \mathcal{A}} V_\xi^\zeta(h_{<t}a_t)$$

$$V_\xi^\zeta(h_{<t}a_t) \;:=\; \frac{1}{\Gamma_t} \lim_{m\to\infty} \sum_{a_{t+1:m},e_{t:m}} \sum_{i=t}^{m} \gamma_i r_i \prod_{j=t}^{m} \xi(e_j|h_{<j}a_j) \prod_{j=t+1}^{m} \zeta(a_j|h_{<j})$$

Note that the value function is still reflective-oracle computable

With the tie breaking trick, $\pi_S \in \mathcal{P}$

Further Thoughts

- Do reflective oracles produce the minimal solution to the grain of truth problem containing the computable measures?
- It seems unrealistic for all agents to share one oracle – what happens if the prior only includes strategies computable with any reflective oracle?
- In the context of reflective oracle access, can we actually show convergence of (the self-predictive) Self-AIXI to (the Bayes-optimal) AIXI?

Sources

[1] M. Hutter, E. Catt, and D. Quarel, An introduction to universal artificial intelligence, First edition. in Chapman & Hall/CRC Artificial Intelligence and robotics series. Boca Raton: Chapman & Hall/CRC Press, 2024

[2] J. Leike, J. Taylor, and B. Fallenstein, “A formal solution to the grain of truth problem,” in Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, in UAI’16. Arlington, Virginia, USA: AUAI Press, Jun. 2016, pp. 427–436.

[3] B. Fallenstein, N. Soares, and J. Taylor, “Reflective Variants of Solomonoff Induction and AIXI,” in Artificial General Intelligence, J. Bieger, B. Goertzel, and A. Potapov, Eds., Cham: Springer International Publishing, 2015, pp. 60–69. doi: 10.1007/978-3-319-21365-1_7.

[4] E. Kalai and E. Lehrer, “Rational Learning Leads to Nash Equilibrium,” Econometrica, vol. 61, no. 5, pp. 1019–1045, 1993, doi: 10.2307/2951492.

[5] B. Fallenstein, J. Taylor, and P. F. Christiano, “Reflective Oracles: A Foundation for Classical Game Theory.” arXiv, Aug. 17, 2015. doi: 10.48550/arXiv.1508.04145.