Universal Knowledge-Seeking Agents in Stochastic Environments

Laurent Orseau[1] Tor Lattimore[2] Marcus Hutter[2]

[1]AgroParisTech, Paris, France

[2]ANU, Canberra, Australia

Algorithmic Learning Theory, 2013

Science

What is science?

- Oxford dictionary:
  The intellectual and practical activity encompassing the **systematic study of the** *structure and behavior of the physical and natural* **world through observation and experiment.**
- Popper + Occam + Epicurus?
  Falsifiability + simplicity + multiple explanations
- What formalization?

Solomonoff induction

- **Formalization + unification + generalization** of **falsifiability + simplicity + multiple explanations**
- Solomonoff prior:

$$\xi(h) := \sum_{\mu \in \mathcal{M}_U} w_\mu \mu(h)$$

$\mathcal{M}_U$: all computable hypotheses

h: observation history

$$w_\mu = 2^{-K(\mu)}$$

K: Kolmogorov complexity

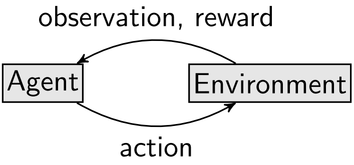
$$\sum_{\mu \in \mathcal{M}_U} w_\mu \leq 1$$

(Kraft inequality)

- Bayes theorem for induction
- Discards inconsistent hypotheses
- **Regret $\leq K(\mu)$ for true environment μ**
- Many good philosophical/logic properties [RH2011]
- **Incomputable** by necessity

Choosing optimal actions

- Induction is not enough:
  observations only, no experiment
- A scientist is active, **must make choices**
  How to choose the **optimal** actions?
- **AIXI** [Hutter2005]
  - **Online** RL setting (no restart)
  - Universal agent based on Solomonoff's prior
  - **Balanced Pareto optimal**



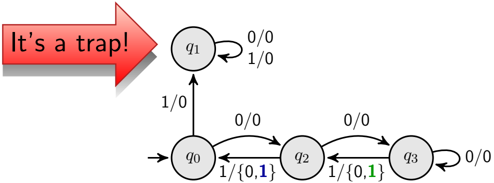
- Almost there, but. . .
  - **Reward-based**, no intrinsic reward function
  - **Exploration issues**

Maximizing prediction accuracy

Intrinsic reward: **maximize prediction accuracy?**

- Bad idea!
  - **May jump into inescapable traps** / kill itself (extreme confirmation bias)

  $\rightarrow$ optimal future prediction for all policies



$\mathcal{M} = \{\mu_1, \mu_2, \mu_3\}$

$\rightarrow$ **Choose actions to maximize long-term expected knowledge**

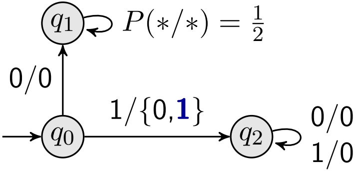
Optimal Scientist

Optimal way to seek knowledge

- Knowledge-seeking agent for **all computable deterministic** environments `[Orseau2011]`
  - Shannon-KSA and Square-KSA
  - Goal: minimize $\xi(h)$
    $\rightarrow$ **Falsifies as many hypotheses as possible**
- Exploration $=$ exploitation
- Convergence to optimal knowledge
  Tends to learn everything it can
- **Avoids traps**

… but fails in stochastic environments

Not resistant to noise



$\mathcal{M} = \{\mu_4, \mu_5\}$

$n\times$ in q_2 loop: $V_{\text{Shannon}} = 1$
$n\times$ in q_1 loop: $V_{\text{Shannon}} = n$

Universal Scientist, v2013

- KL-KSA, based on Kullback-Leibler divergence

$$V^\pi := \sum_{\mu \in \mathcal{M}} 2^{-K(\mu)} KL^\pi(\mu || \xi)$$
$$\pi^* := \arg\max_\pi V^\pi$$

- Maximize the **expected divergence** between

  each individual possible **environment** and
  the **agent's knowledge** of the world.

→ Choose actions that **maximize expected information gain**

- **Time consistency:**
  Choosing π^* at $t = 0$ and following it after history h
  same as choosing π^* after history h.

Convergence

Theorems:

- **On-policy prediction**
  - Learns to predict accurately the future history
  - (True for all policies)

(main theorem)

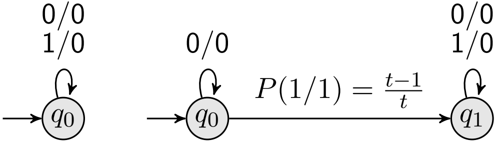
- **On-policy learning, off-policy prediction**
  - Learns to predict if would follow *any* policy
  - Reason:
    π^* **outcomes are the most difficult to predict**

Noise and traps

- Non-informative policy π:
  Outcomes have equal probabilities for all (consistent) environments
  $\rightarrow KL^{\pi} = 0$
- Noise: non-informative π with stochastic outcomes
  $\rightarrow V^{\pi} = 0$
  $\rightarrow$ **KL-KSA resistant to noise**
- Trap: all policies are non-informative
  $\rightarrow \forall \pi V^{\pi} = 0$
  $\rightarrow$ **KL-KSA avoids traps**

KL-KSA, undiscounted: Issues

- **Non-existence of the value** for $\mathcal{M} = \mathcal{M}_U$
  KL-entropy(ξ) $\geq \sum_x 2^{-K(x)} K(x) = \infty$
  $\rightarrow V^{\pi^*} = \infty$
- **Non-existence of the optimal policy**
  Even if value existed



μ_6 μ_7

  - Less clear if $\mathcal{M} = \mathcal{M}_U \ldots$

Solution 1: Horizon function

- Weights γ_t each time step (finite sum)
- Need to define discounted KL_γ
- **Ensures existence of value + policy**
- But not appealing
  - **Myopic**
  - **No fundamentally justified choice**
  - **Infinite dimension vector**

Solution 2: Approximations

- ϵ-biased prior: $w_\mu = 2^{-(1+\epsilon)K(\mu)}$
  - **Existence of the value**
    (finite entropy)
  - But **loses dominance property**
- δ-optimal policy
  - **Existence of the (near-)optimal policy**
  - But **may stop exploring at some point**
- **Only 2 scalar parameters**

Conclusion

What is science?

Choose actions to maximize long-term expected knowledge

- **First formal definition of the optimal scientific process** for all computable stochastic environments
- **Still some annoying parameters**
  - Horizon function, $\epsilon-$biased prior $+$ $\delta-$optimal policy
  - Reference machine
- Rate of convergence?
- **How to be more convincing?**
  - How to *prove* this defines (or not) science? What mathematical properties are required?

Bibliography

- [Hutter2005] M. Hutter, Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability, Springer, 2005.
- [RH2011] S. Rathmanner and M. Hutter, *A Philosophical Treatise of Universal Induction*, Entropy (13) 6, 1076–1136, 2011.
- [Orseau2011] L. Orseau, *Universal Knowledge-Seeking Agents*. ALT (6925), 353–367, 2011.