Universal Reinforcement Learning Algorithms: Survey and Experiments

John Aslanides[†], Jan Leike[‡], Marcus Hutter[†]
{john.aslanides,marcus.hutter}@anu.edu.au, leike@google.com

† Australian National University
‡ Future of Humanity Institute

IJCAI 2017

August 24, 2017

Overview

- Introduction
- Algorithms
- Experiments

Motivation

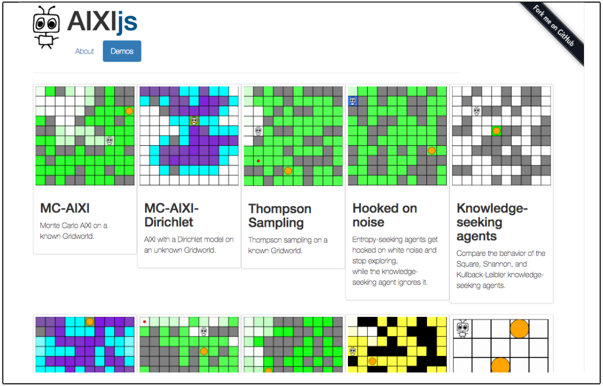
- **Universal RL** (URL): *making very weak assumptions about its environment, what can an agent achieve, in principle?*
  - What is intelligent behavior?
  - What is a useful optimality criterion?
- Theoretically studied, but few to no experiments/reference implementations to date

Motivation

- **Universal RL** (URL): *making very weak assumptions about its environment, what can an agent achieve, in principle?*
    - What is intelligent behavior?
    - What is a useful optimality criterion?
- Theoretically studied, but few to no experiments/reference implementations to date
- Contribution: experiments, along with open-source demo platform for several URL algorithms.

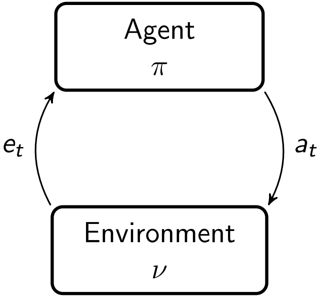
Demo

Online demo: http://aslanides.io/aixijs



Agent-environment model

- Environment class: **POMDPs** (possibly non-ergodic)
- Percepts ($\neq$ states) are (**observation**, **reward**) pairs $e_k = (o_k, r_k)$
- Interact to generate a **history** $h_t := a_1 e_1 a_2 e_2 \dots a_t e_t$.



AIξ (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable **model class** $\mathcal{M}$:

$$\xi\left(e_t|h_t\right) \doteq \sum_{\nu\in\mathcal{M}} w\left(\nu|h_t\right)\nu\left(e_t|h_t\right)$$

AIξ (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable **model class** $\mathcal{M}$:

$$\xi\left(e_t|h_t\right) \doteq \sum_{\nu\in\mathcal{M}} w\left(\nu|h_t\right)\nu\left(e_t|h_t\right)$$

- The **Bayes-optimal policy** (AIξ; Hutter, 2005) is

$$a_t = \underbrace{\arg\max_{a_t}\sum_{e_t}\cdots\max_{a_m}\sum_{e_m}}_{\text{expectimax search}}\underbrace{\sum_{k=t}^{m}\gamma_k u\left(h_k\right)}_{\text{return}}\underbrace{\prod_{j=t}^{k}\xi\left(e_j|h_t\right)}_{\text{environment model}}.$$

AIξ (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable **model class** $\mathcal{M}$:

$$\xi\left(e_t|h_t\right) \doteq \sum_{\nu\in\mathcal{M}} w\left(\nu|h_t\right)\nu\left(e_t|h_t\right)$$

- The **Bayes-optimal policy** (AIξ; Hutter, 2005) is

$$a_t = \underbrace{\arg\max_{a_t}\sum_{e_t}\cdots\max_{a_m}\sum_{e_m}}_{\text{expectimax search}}\underbrace{\sum_{k=t}^{m}\gamma_k u\left(h_k\right)}_{\text{return}}\underbrace{\prod_{j=t}^{k}\xi\left(e_j|h_t\right)}_{\text{environment model}}.$$

- In practice:
  - Forward planning by MCTS
  - Use manageable model class $\mathcal{M}$

AIξ (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable **model class** $\mathcal{M}$:

$$\xi\left(e_t | h_t\right) \doteq \sum_{\nu \in \mathcal{M}} w\left(\nu | h_t\right) \nu\left(e_t | h_t\right)$$

- The **Bayes-optimal policy** (AIξ; Hutter, 2005) is

$$a_t = \underbrace{\arg\max_{a_t} \sum_{e_t} \cdots \max_{a_m} \sum_{e_m}}_{\text{expectimax search}} \underbrace{\sum_{k=t}^{m} \gamma_k u\left(h_k\right)}_{\text{return}} \underbrace{\prod_{j=t}^{k} \xi\left(e_j | h_t\right)}_{\text{environment model}} .$$

- In practice:
  - Forward planning by MCTS
  - Use manageable model class $\mathcal{M}$

- [Video]

Issues

Problems:

- Not asymptotically optimal (Orseau, 2010)

Issues

Problems:

- Not asymptotically optimal (Orseau, 2010)
- Won't overcome the bias of bad priors, *c.f.* supervised learning (Leike & Hutter, 2015)

Knowledge-seeking agents (Orseau, 2013)

- **Utility agents** are intrinsically motivated, and don't need an extrinsic reward signal.
- Knowledge-seeking agent (KSA) – motivated to reduce uncertainty
- No exploration/exploitation tradeoff

Agent	**Utility function**	**Description**
AIξ	r	Reward
Square-KSA	$-\xi$	Entropy
Shannon-KSA	$-\log[\xi]$	Entropy
Kullback-Leibler-KSA	$-\Delta\text{Ent}[w(\cdot)]$	Information gain

Knowledge-seeking agents (Orseau, 2013)

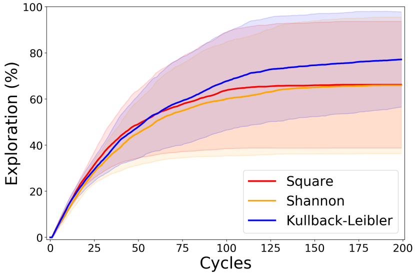
- **Utility agents** are intrinsically motivated, and don't need an extrinsic reward signal.
- Knowledge-seeking agent (KSA) – motivated to reduce uncertainty
- No exploration/exploitation tradeoff

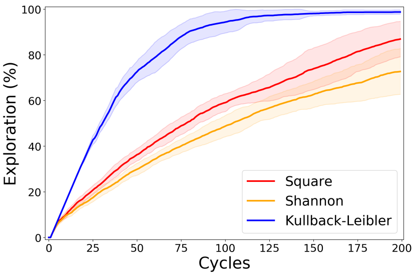
Agent	**Utility function**	**Description**
AIξ	r	Reward
Square-KSA	$-\xi$	Entropy
Shannon-KSA	$-\log[\xi]$	Entropy
Kullback-Leibler-KSA	$-\Delta\mathrm{Ent}[w(\cdot)]$	Information gain

- [Video]

Experiments

- Qualitative behavior is highly model-sensitive
- KL-KSA outperforms entropy-seeking in stochastic environments





Outlook

- Open-source online JavaScript demo: https://aslanides.io/aixijs
- Used to run experiments for another IJCAI paper (Reinforcement Learning with a Corrupted Reward Channel, Everitt et al. 2017)
- Come and talk to me at the ANU booth downstairs :)