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
Introduction

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.
Online demo: http://aslanides.io/aixijs
Agent-environment model

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

![Diagram of Agent-environment model](image)
$\text{AI}\xi$ (Hutter, 2005)

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

$$\xi(e_t|h_t) = \sum_{\nu \in \mathcal{M}} w(\nu|h_t) \nu(e_t|h_t)$$
Alξ (Hutter, 2005)

- Non-parametric Bayesian mixture over some countable model class $\mathcal{M}$:
  $$\xi(e_t | h_t) = \sum_{\nu \in \mathcal{M}} w(\nu | h_t) \nu(e_t | h_t)$$

- The Bayes-optimal policy (Alξ; Hutter, 2005) is
  $$a_t = \arg\max_{a_t} \sum_{e_t} \cdots \max_{a_m} \sum_{e_m} \sum_{k=t}^{m} \gamma_k u(h_k) \prod_{j=t}^{k} \xi(e_j | h_t)$$
  
  expectimax search \hspace{1cm} return \hspace{1cm} environment model
Al$\xi$ (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable model class $\mathcal{M}$:
  \[
  \xi(e_t|h_t) = \sum_{\nu \in \mathcal{M}} w(\nu|h_t) \nu(e_t|h_t)
  \]

- The **Bayes-optimal policy** (Al$\xi$; Hutter, 2005) is
  \[
  a_t = \arg\max_{a_t} \sum_{e_t} \cdots \max_{a_m} \sum_{e_m} \sum_{k=t}^{m} \gamma_k u(h_k) \prod_{j=t}^{k} \xi(e_j|h_t)
  \]
  
  In practice:
  - Forward planning by MCTS
  - Use manageable model class $\mathcal{M}$
Al\(\xi\) (Hutter, 2005)

- Non-parametric **Bayesian mixture** over some countable model class \(\mathcal{M}\):
  \[
  \xi(e_t|h_t) = \sum_{\nu \in \mathcal{M}} w(\nu|h_t) \nu(e_t|h_t)
  \]

- The **Bayes-optimal policy** (Al\(\xi\); Hutter, 2005) is
  \[
  a_t = \arg \max_a \sum_{e_t} \cdots \max_a \sum_{e_m} \sum_{k=t}^{m} \gamma_k u(h_k) \prod_{j=t}^k \xi(e_j|h_t)
  \]
  
  - **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

<table>
<thead>
<tr>
<th>Agent</th>
<th>Utility function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI$\xi$</td>
<td>$r$</td>
<td>Reward</td>
</tr>
<tr>
<td>Square-KSA</td>
<td>$-\xi$</td>
<td>Entropy</td>
</tr>
<tr>
<td>Shannon-KSA</td>
<td>$-\log[\xi]$</td>
<td>Entropy</td>
</tr>
<tr>
<td>Kullback-Leibler-KSA</td>
<td>$-\Delta\text{Ent}[w(\cdot)]$</td>
<td>Information gain</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Agent</th>
<th>Utility function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI $\xi$</td>
<td>$r$</td>
<td>Reward</td>
</tr>
<tr>
<td>Square-KSA</td>
<td>$-\xi$</td>
<td>Entropy</td>
</tr>
<tr>
<td>Shannon-KSA</td>
<td>$-\log [\xi]$</td>
<td>Entropy</td>
</tr>
<tr>
<td>Kullback-Leibler-KSA</td>
<td>$-\Delta \text{Ent} [w(\cdot)]$</td>
<td>Information gain</td>
</tr>
</tbody>
</table>

[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 :)