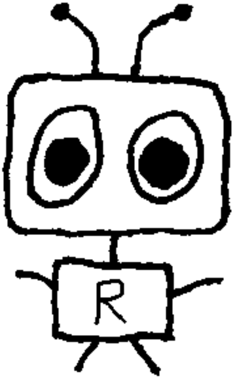
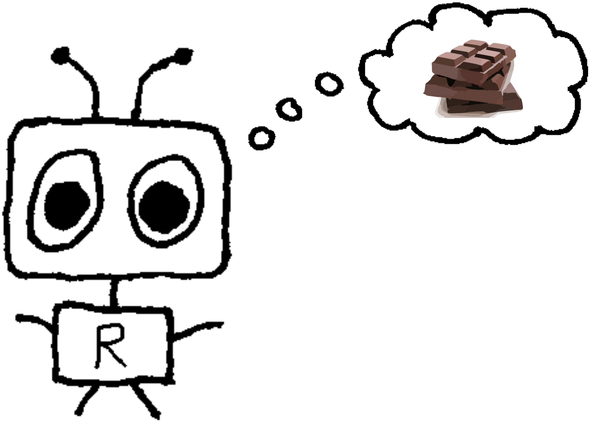
Bayesian Reinforcement Learning

Jan Leike

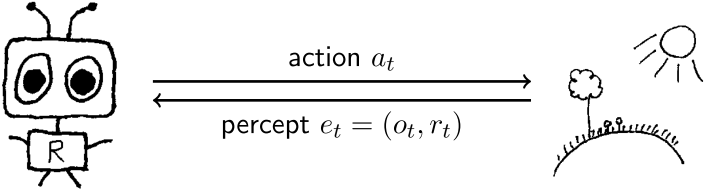
Australian National University

21 December 2015

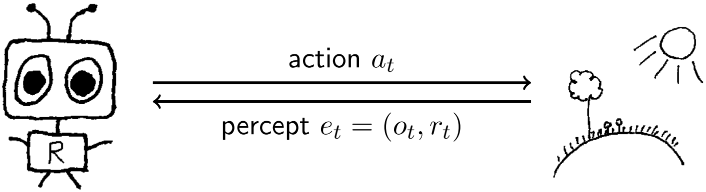

R


R

Reinforcement Learning

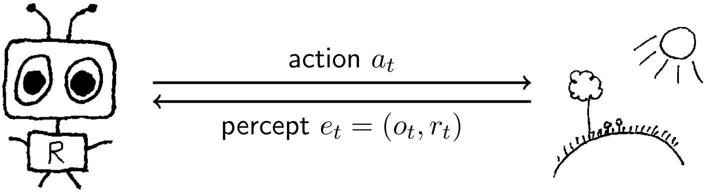


Reinforcement Learning



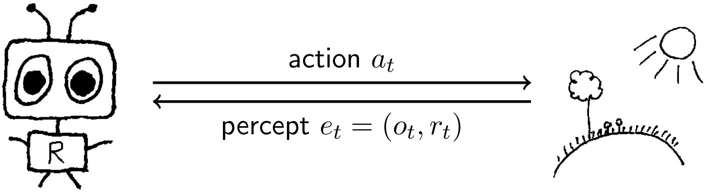
history $\quad$ $æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$

Reinforcement Learning



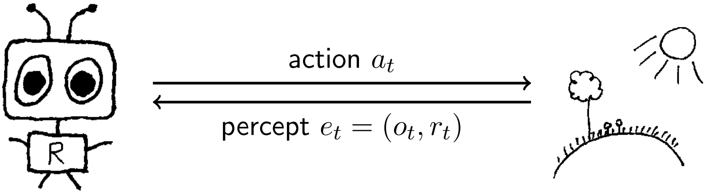
history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
policy	π : Histories $\rightsquigarrow$ Actions

Reinforcement Learning



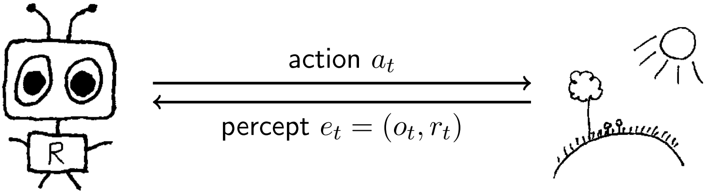
history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
policy	π : Histories $\rightsquigarrow$ Actions
environment	ν : Histories $\times$ Actions $\rightsquigarrow$ Percepts

Reinforcement Learning



history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
policy	π : Histories $\rightsquigarrow$ Actions
environment	ν : Histories $\times$ Actions $\rightsquigarrow$ Percepts
true environment	μ

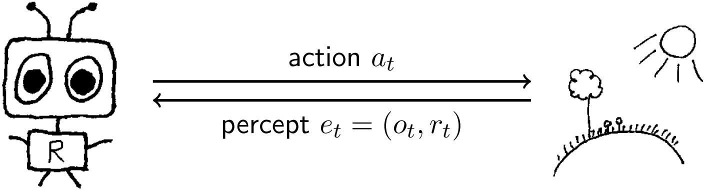
Reinforcement Learning



history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
policy	π : Histories $\rightsquigarrow$ Actions
environment	ν : Histories $\times$ Actions $\rightsquigarrow$ Percepts
true environment	μ

Goal: maximize $\sum_{t=1}^{\infty} \gamma(t) r_t$
where $\gamma : \mathbb{N} \to [0, 1]$ is a discount function with $\sum_{t=1}^{\infty} \gamma(t) < \infty$

Reinforcement Learning



history	$æ_{<t} := a_1 e_1 a_2 e_2 \dots a_{t-1} e_{t-1}$
policy	π : Histories $\rightsquigarrow$ Actions
environment	ν : Histories $\times$ Actions $\rightsquigarrow$ Percepts
true environment	μ

Goal: maximize $\sum_{t=1}^{\infty} \gamma(t) r_t$
where $\gamma : \mathbb{N} \to [0,1]$ is a discount function with $\sum_{t=1}^{\infty} \gamma(t) < \infty$

Assume: $0 \leq r_t \leq 1$

Value Functions

Value Functions

Value of policy π in environment ν:

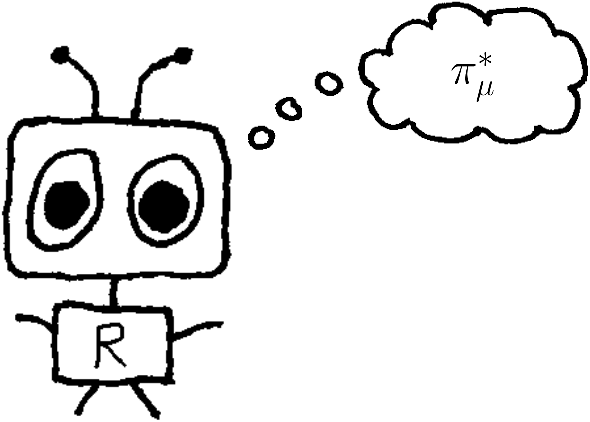
$$V^\pi_\nu(æ_{<t}) := \frac{1}{\Gamma_t}\mathbb{E}^\pi_\nu\left[\sum_{k=t}^{\infty}\gamma(k)r_k \,\middle|\, æ_{<t}\right]$$

Value Functions

Value of policy π in environment ν:

$$V_\nu^\pi(æ_{<t}) := \frac{1}{\Gamma_t}\mathbb{E}_\nu^\pi\left[\sum_{k=t}^{\infty}\gamma(k)r_k \,\middle|\, æ_{<t}\right]$$

Optimal value: $V_\nu^* := \sup_\pi V_\nu^\pi$
ν-optimal policy: $\pi_\nu^* := \arg\max_\pi V_\nu^\pi$


R
π^*_μ

AIXI[2]

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

AIXI[2]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

AIXI[2]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

AIXI[2]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$
- Bayesian mixture

$$\xi := \sum_{\nu \in \mathcal{M}} w(\nu)\nu$$

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

AIXI[2]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$
- Bayesian mixture

$$\xi := \sum_{\nu \in \mathcal{M}} w(\nu)\nu$$

Solomonoff:[1] $w(\nu) := 2^{-K(\nu)}$,
$K(\nu) :=$ length of the shortest description of ν

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.

AIXI[2]

- countable set of environments $\mathcal{M} = \{\nu_1, \nu_2, \dots\}$
- prior $w : \mathcal{M} \to [0, 1]$
- Bayesian mixture

$$\xi := \sum_{\nu \in \mathcal{M}} w(\nu)\nu$$

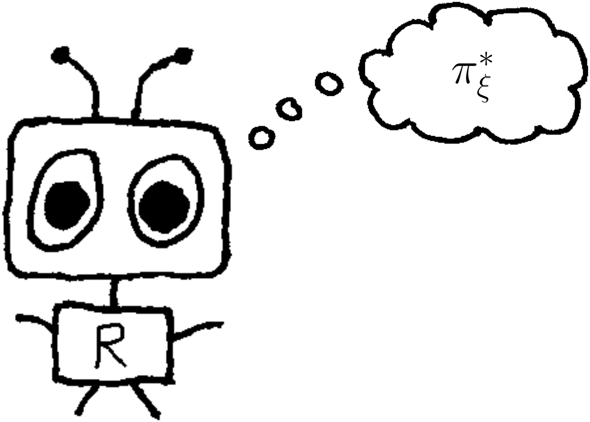
Solomonoff:[1] $w(\nu) := 2^{-K(\nu)}$,
$K(\nu) :=$ length of the shortest description of ν

AIXI is the Bayes-optimal agent with a Solomonoff prior

$$\pi^*_\xi := \arg\max_\pi V^\pi_\xi$$

[1] Ray Solomonoff. "A Formal Theory of Inductive Inference. Parts 1 and 2". In: *Information and Control* 7.1 (1964), pages.

[2] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2005.


R
π^*_ξ

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V_\nu^\pi$$

[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V^\pi_\nu = V^\pi_\xi$$

[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

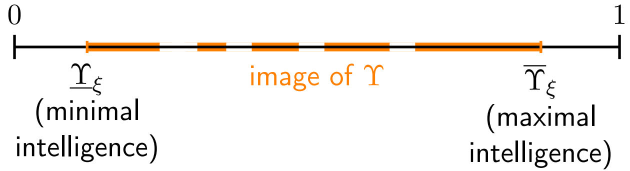
$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V^\pi_\nu = V^\pi_\xi$$

[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V^\pi_\nu = V^\pi_\xi$$

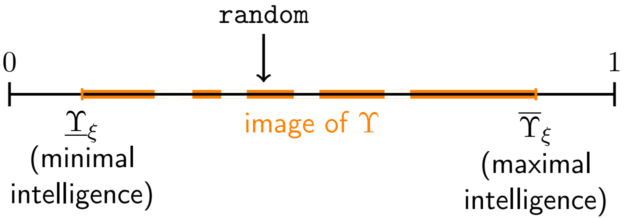


[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V^\pi_\nu = V^\pi_\xi$$

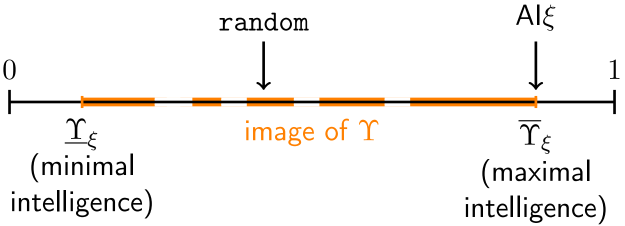


[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.

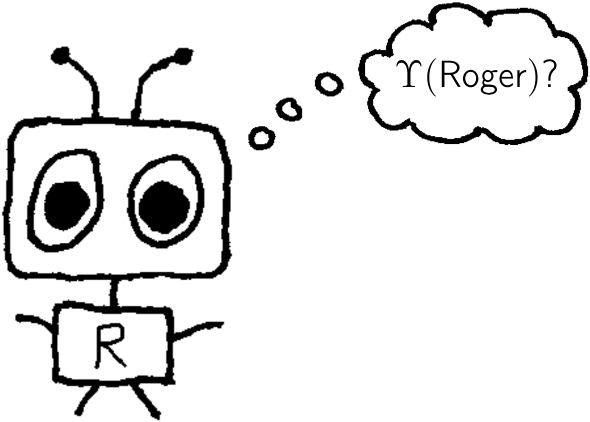
Legg-Hutter Intelligence[3]

Intelligence measures an agent's ability to achieve goals in a wide range of environments.

$$\Upsilon_\xi(\pi) := \sum_\nu w(\nu) V^\pi_\nu = V^\pi_\xi$$



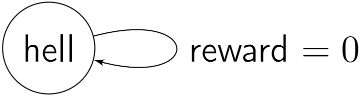
[3] Shane Legg and Marcus Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds & Machines* 17.4 (2007), pp. 391–444.



Υ(Roger)?

Hell

Hell



The Dogmatic Prior[4]

[4] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[4]

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

[4] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[4]

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

if not acting according to π_{Lazy},
go to hell with high probability

[4] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

The Dogmatic Prior[4]

Policy π_{Lazy}:

```
while (true) { do_nothing(); }
```

Dogmatic prior ξ':

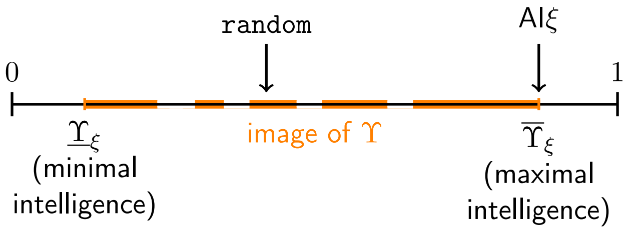
if not acting according to π_{Lazy},
go to hell with high probability

Theorem

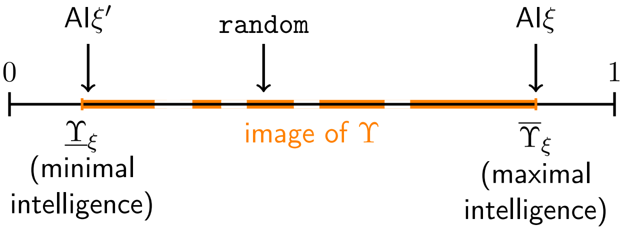
AIξ' acts according to π_{Lazy} as long as $V_{\xi}^{\pi_{Lazy}}(æ_{<t}) > \varepsilon > 0$ (future expected reward does not get close to 0).

[4] Jan Leike and Marcus Hutter. "Bad Universal Priors and Notions of Optimality". In: *Conference on Learning Theory*. 2015, pp. 1244–1259.

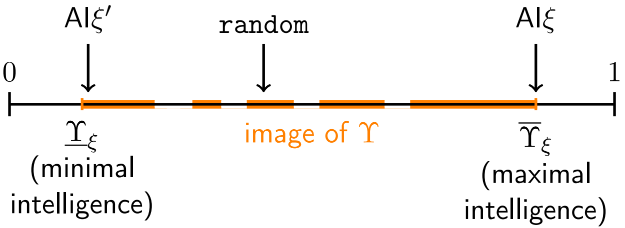
Consequences for Intelligence



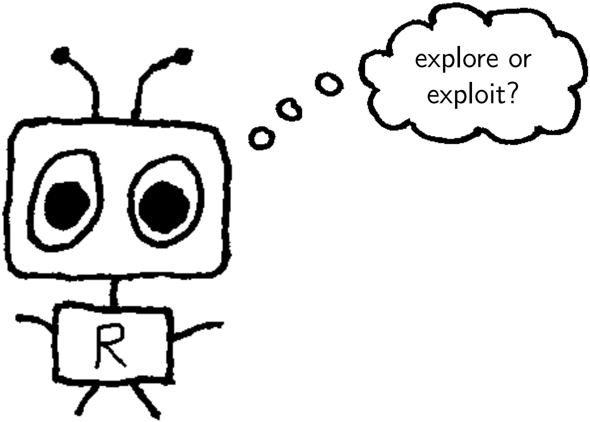
Consequences for Intelligence



Consequences for Intelligence



$\Longrightarrow$ **Legg-Hutter intelligence is highly subjective**


explore or
exploit?
R

Asymptotic Optimality

π is *asymptotically optimal* iff

$$V^*_\mu(æ_{<t}) - V^\pi_\mu(æ_{<t}) \to 0 \text{ as } t \to \infty$$

[5] Laurent Orseau. "Asymptotic Non-Learnability of Universal Agents with Computable Horizon Functions". In: *Theoretical Computer Science* 473 (2013), pp. 149–156.

Asymptotic Optimality

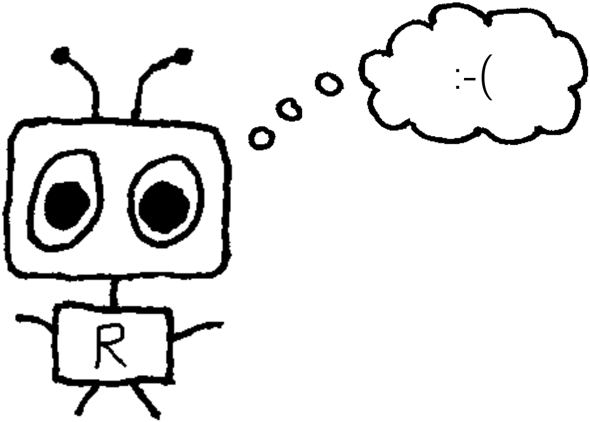
π is *asymptotically optimal* iff

$$V^*_\mu(æ_{<t}) - V^\pi_\mu(æ_{<t}) \to 0 \text{ as } t \to \infty$$

Theorem

AIXI is not asymptotically optimal.[5]

[5] Laurent Orseau. "Asymptotic Non-Learnability of Universal Agents with Computable Horizon Functions". In: *Theoretical Computer Science* 473 (2013), pp. 149–156.


R
:-(

Knowledge-Seeking

- $m \in \mathbb{N}$ is the horizon

[6] Laurent Orseau, Tor Lattimore, and Marcus Hutter. "Universal Knowledge-Seeking Agents for Stochastic Environments". In: *Algorithmic Learning Theory*. Springer, 2013, pp. 158–172.

Knowledge-Seeking

- $m \in \mathbb{N}$ is the horizon

Information-seeking policy[6]

$$\pi_I^* := \arg\max_\pi \mathbb{E}_{\nu \sim w(\,\cdot \mid æ_{<t})}[\mathrm{KL}_{1:m}(\nu^\pi, \xi^\pi)]$$

[6] Laurent Orseau, Tor Lattimore, and Marcus Hutter. "Universal Knowledge-Seeking Agents for Stochastic Environments". In: *Algorithmic Learning Theory*. Springer, 2013, pp. 158–172.

Knowledge-Seeking

- $m \in \mathbb{N}$ is the horizon

Information-seeking policy[6]

$$\pi_I^* := \arg\max_\pi \mathbb{E}_{\nu \sim w(\,\cdot \mid æ_{<t})}[\mathrm{KL}_{1:m}(\nu^\pi, \xi^\pi)]$$
$$= \arg\max_\pi \mathbb{E}_\xi^\pi[\mathrm{Ent}(w(\,\cdot \mid æ_{<t})) - \mathrm{Ent}(w(\,\cdot \mid æ_{1:m}))]$$

[6] Laurent Orseau, Tor Lattimore, and Marcus Hutter. "Universal Knowledge-Seeking Agents for Stochastic Environments". In: *Algorithmic Learning Theory*. Springer, 2013, pp. 158–172.

Knowledge-Seeking

- $m \in \mathbb{N}$ is the horizon

Information-seeking policy[6]

$$\pi_I^* := \arg\max_\pi \mathbb{E}_{\nu \sim w(\,\cdot \mid æ_{<t})}[\mathrm{KL}_{1:m}(\nu^\pi, \xi^\pi)]$$
$$= \arg\max_\pi \mathbb{E}_\xi^\pi[\mathrm{Ent}(w(\,\cdot \mid æ_{<t})) - \mathrm{Ent}(w(\,\cdot \mid æ_{1:m}))]$$

Effective horizon:

$$H_t(\varepsilon) := \min \left\{ k \,\middle|\, \frac{\sum_{i=t+k}^\infty \gamma(i)}{\sum_{i=t}^\infty \gamma(i)} \leq \varepsilon \right\}$$

[6] Laurent Orseau, Tor Lattimore, and Marcus Hutter. "Universal Knowledge-Seeking Agents for Stochastic Environments". In: *Algorithmic Learning Theory*. Springer, 2013, pp. 158–172.

BayesExp[7]

BayesExp:

> $\mathit{if}\ \mathbb{E}_{\nu \sim w(\,\cdot \mid æ_{<t})}[\mathrm{KL}_{1:m}(\nu^\pi, \xi^\pi)] > \varepsilon_t$
> *then execute* π^*_I *for* $H_t(\varepsilon_t)$ *steps*
> *else execute* π^*_ξ *for* 1 *step*

with $\varepsilon_t \to 0$ as $t \to \infty$

[7] Tor Lattimore. "Theory of General Reinforcement Learning". PhD thesis. Australian National University, 2013, Chapter 5.

BayesExp[7]

BayesExp:

> *if* $\mathbb{E}_{\nu \sim w(\,\cdot \mid æ_{<t})}[\mathrm{KL}_{1:m}(\nu^\pi, \xi^\pi)] > \varepsilon_t$
> *then execute* π_I^* *for* $H_t(\varepsilon_t)$ *steps*
> *else execute* π_ξ^* *for* 1 *step*

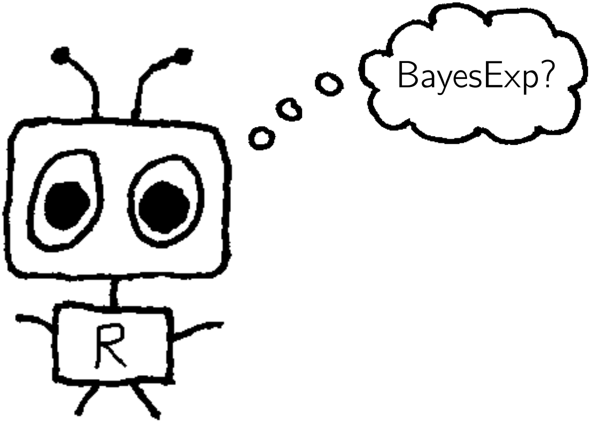
with $\varepsilon_t \to 0$ as $t \to \infty$

Theorem

BayesExp is asymptotically optimal:

$$\frac{1}{n}\sum_{t=1}^{n}\big(V_\mu^*(æ_{<t}) - V_\mu^\pi(æ_{<t})\big) \to 0 \text{ as } t \to \infty \ \mu\text{-almost surely}$$

[7] Tor Lattimore. "Theory of General Reinforcement Learning". PhD thesis. Australian National University, 2013, Chapter 5.


BayesExp?

Summary

Summary

- For Bayesian RL the prior matters

Summary

- For Bayesian RL the prior matters
- Bad priors are bad

Summary

- For Bayesian RL the prior matters
- Bad priors are bad
- Good priors are not asymptotically optimal

Summary

- For Bayesian RL the prior matters
- Bad priors are bad
- Good priors are not asymptotically optimal
- Asymptotic optimality needs more exploration

Summary

- For Bayesian RL the prior matters
- Bad priors are bad
- Good priors are not asymptotically optimal
- Asymptotic optimality needs more exploration
- Do we want asymptotic optimality?