A Model-Free Universal AI

Yegon Kim and Juho Lee

Presenter: Yegon Kim / Date: 2026.03.23

KAIST

Problem

- Model-based vs model-free
- AIXI
- Can a model-free agent be optimal in general environments?

Related Work

- Feature RL
- Optimal Direct Policy Search
- Self-AIXI

Background

Policy π interacts with environment ν to create history $h_{1:t}$

$$h_{1:t} = a_1 e_1 \; \dots \; a_t e_t$$

$$\nu^\pi(h_{1:t}) := \pi(a_1)\,\nu(e_1|a_1) \cdots \pi(a_t|h_{<t})\,\nu(e_t|h_{<t}a_t)$$

Percept e_t contains observation o_t and reward r_t

Background

Given discount factor $0 < \gamma < 1$,

$$R_t := (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

Value functions are expected returns:

$$V_\nu^\pi(h_{<t}) := \mathbb{E}_\nu^\pi \left[R_t \mid h_{<t} \right]$$

$$Q_\nu^\pi(h_{<t}, a_t) := \mathbb{E}_\nu^\pi \left[R_t \mid h_{<t} a_t \right]$$

Optimal policy π_ν^* is a policy with maximum possible value.

AIXI

Mixture environment ξ, given $w(\nu)$ over $\nu \in \mathcal{M}$:

$$\xi(e_t|h_{<t}a_t) := \sum_{\nu \in \mathcal{M}} w(\nu|h_{<t})\nu(e_t|h_{<t}a_t) \quad \textit{with posterior} \quad w(\nu|h_{1:t}) := w(\nu|h_{<t}) \frac{\nu(e_t|h_{<t}a_t)}{\xi(e_t|h_{<t}a_t)}$$

AIXI

AIXI is the optimal policy π^*_ξ in ξ. It is *Bayes-optimal* by design.

A straightforward (finite horizon) implementation:

$$a_t = \operatorname*{argmax}_{a_t} \sum_{e_t} \ldots \max_{a_m} \sum_{e_m} (r_t + \ldots \gamma^{t-m} r_m)\, \xi(e_{t:m} \parallel a_{t:m} \mid h_{<t})$$

Universal AI with Q-Induction (AIQI)

Overview

At every time step,

1. Predict (distribution of) returns given history $h_{<t}$ and action a_t

2. Pick action a_t with largest expected return, a.k.a. Q-value

   $+$ ε-greedy exploration

Universal AI with Q-Induction (AIQI)

H-step return

$$R_{t,H} := (1-\gamma)\sum_{k=0}^{H-1}\gamma^k r_{t+k}$$

Discretized (H-step) return

$$z_t := \frac{\lfloor M R_{t,H} \rfloor}{M}, \quad z_t \in \mathcal{Z} := \left\{0, \frac{1}{M}, \dots, \frac{M-1}{M}\right\}$$

Universal AI with Q-Induction (AIQI)

How do we predict z_t ?

Universal AI with Q-Induction (AIQI)

How do we predict z_t ?

Sequence prediction with

$$\dots a_{t-3} z_{t-3} e_{t-3} a_{t-2} z_{t-2} e_{t-2} a_{t-1} z_{t-1} e_{t-1} a_t$$

doesn't work. (Why?)

Universal AI with Q-Induction (AIQI)

How do we predict z_t ?

Sequence prediction with

$$\dots a_{t-3} z_{t-3} e_{t-3} a_{t-2} z_{t-2} e_{t-2} a_{t-1} z_{t-1} e_{t-1} a_t$$

doesn't work. (Why?)

Instead we use the periodically augmented history

$$\dots a_{t-2N} z_{t-2N} e_{t-2N} \dots a_{t-N} z_{t-N} e_{t-N} \dots a_{t-1} e_{t-1} a_t$$

Universal AI with Q-Induction (AIQI)

A phase n return-predictor ϕ maps phase n augmented history

$$a_1 e_1 \ldots a_n z_n e_n \ldots a_{n+kN}$$

to distribution over returns $\tilde{z}_{n+kN}$.

Given a hypothesis class $\mathscr{P}_n$ of phase n return-predictors and a prior ω_n over them, we can define the phase n mixture return-predictor ψ_n.

Universal AI with Q-Induction (AIQI)

We thus have N mixture *return-predictors* ψ_n.

We can define a single, unified return-predictor ψ

$$\psi(\tilde{z}_t \mid h_{<t}a_t) := \psi_n(\tilde{z}_t \mid \mathrm{aug}_n(h_{<t})a_t)$$

with which we make the Q-value estimate

$$\hat{Q}(h_{<t}, a_t) = \sum_{\tilde{z}_t \in \mathcal{Z}} \tilde{z}_t \cdot \psi(\tilde{z}_t \mid h_{<t}a_t)$$

Universal AI with Q-Induction (AIQI)

AIQI chooses largest Q-value action, with random exploration τ

$$\hat{\pi}(a \mid h_{<t}) := (1-\tau)\mathbb{1}\left[a = a^*\right] + \tau/|\mathcal{A}|, \quad \textit{where } a^* = \arg\max_{a_t} \hat{Q}(h_{<t}, a_t)$$

Intuitively, we need the exploration so that return-predictor becomes accurate for all actions a, not just for $a*$.

Universal AI with Q-Induction (AIQI)

Parameters:

- Horizon H

- Return discretization level M

- Period N

- Exploration τ

Theoretical Results

<u>Grain of truth</u>

Recall that ψ_n is a mixture of return-predictors $\phi \in \mathscr{P}_n$.

$\mathscr{P}_n$ should contain the true return-predictor ϕ^* induced by AIQI policy $\hat{\pi}$, which depends on $\mathscr{P}_n$.

Theoretical Results

Policy π is strong asymptotically ε-optimal in $\mathscr{M}$ if:

For all $\nu \in \mathscr{M}$,

$$\limsup_{t\to\infty} V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t}) \leq \varepsilon, \quad \nu^\pi\text{-}a.s.$$

With the right parameters H, M, N, τ, ψ, AIQI is strong asymptotically ε-optimal. (Theorem 4.6)

$$\tau \leq \frac{\varepsilon(1-\gamma)}{10},\ M \geq \frac{10}{\varepsilon(1-\gamma)},\ H = H(\eta) \text{ with } \eta \leq \frac{\varepsilon(1-\gamma)}{10},\ N \geq H + \log_\gamma \frac{\varepsilon}{5}$$

Proof

With Blackwell-Dubins theorem, we can show that the mixture return-predictor converges to the *true* return-predictor. (Lemma 4.2; grain of truth is used here)

$$\sum_{\tilde{z}_t \in \mathcal{Z}} |\psi(\tilde{z}_t \mid h_{<t}\, a_t) - \nu^{\pi}(\tilde{z}_t \mid h_{<t}\, a_t)| \qquad \text{becomes small}$$

Proof

With Blackwell-Dubins theorem, we can show that the mixture return-predictor converges to the *true* return-predictor. (Lemma 4.2; grain of truth is used here)

Good return-predictor $\Longrightarrow$ good value prediction (Lemma 4.3)

$$|\hat{Q}(h_{<t}a_t) - Q^{\pi}_{\nu}(h_{<t}a_t)| \qquad \text{becomes small}$$

Proof

With Blackwell-Dubins theorem, we can show that the mixture return-predictor converges to the *true* return-predictor. (Lemma 4.2; grain of truth is used here)

Good return-predictor $\implies$ good value prediction (Lemma 4.3)

Good value prediction $\implies$ "one-step optimal" action choice (Lemma 4.4)

$$\delta_1(h_{<t}) := \max_a Q_\nu^\pi(h_{<t}, a) - V_\nu^\pi(h_{<t}) \quad \text{becomes small}$$

Proof

With Blackwell-Dubins theorem, we can show that the mixture return-predictor converges to the *true* return-predictor. (Lemma 4.2; grain of truth is used here)

Good return-predictor $\Longrightarrow$ good value prediction (Lemma 4.3)

Good value prediction $\Longrightarrow$ "one-step optimal" action choice (Lemma 4.4)

One-step optimal $\Longrightarrow$ globally optimal (Lemma 4.5)

$$\delta_\infty(h_{<t}) := V^*_\nu(h_{<t}) - V^\pi_\nu(h_{<t}) \quad \text{becomes small}$$

Theoretical Results

AIQI is asymptotically ε-optimal in ξ. (Theorem 4.8)

$$\limsup_{t\to\infty} V_\xi^*(h_{<t}) - V_\xi^\pi(h_{<t}) \leq \varepsilon \ \ \textit{holds both } \xi^\pi\textit{-a.s. and, for all } \nu \in \mathcal{M}, \ \nu^\pi\textit{-a.s.}$$

Also, any infinite repeated game of AIQIs converges to ε-Nash equilibrium.

Theoretical Results

AIQI is similar to on-policy Monte Carlo control.

AIQI does *not* do well on off-policy histories, unlike AIXI.

Theoretical Results

Self-optimizing w.r.t. historic policy π', environment class $\mathcal{M}$

For any $\nu \in \mathcal{M}$,

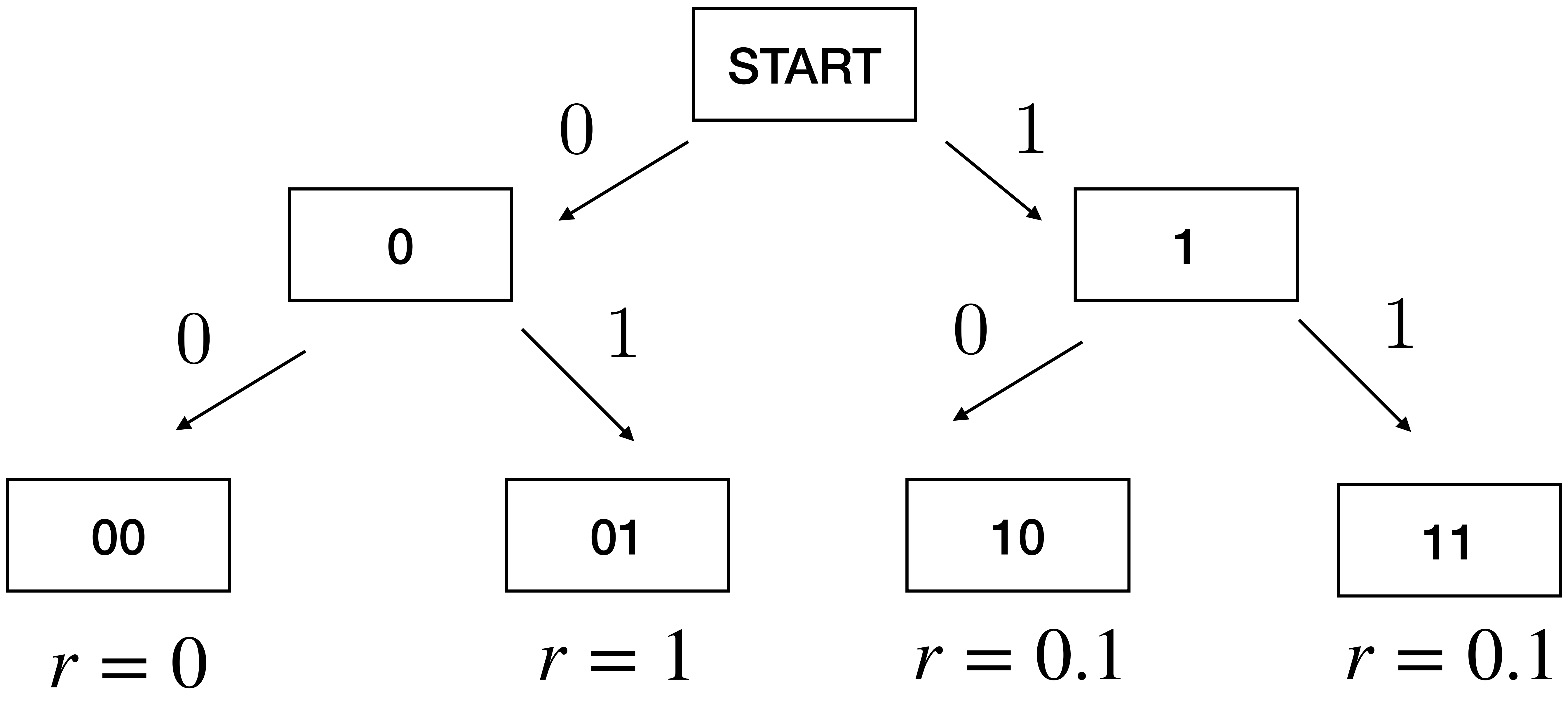
$$\lim_{t \to \infty} V_\nu^*(h_{<t}) - V_\nu^{\bar{\pi}}(h_{<t}) = 0, \quad \nu^{\pi'}\text{-}a.s.$$

AIXI is self-optimizing w.r.t. π', $\mathcal{M}$ if there is such a policy.

Theoretical Results

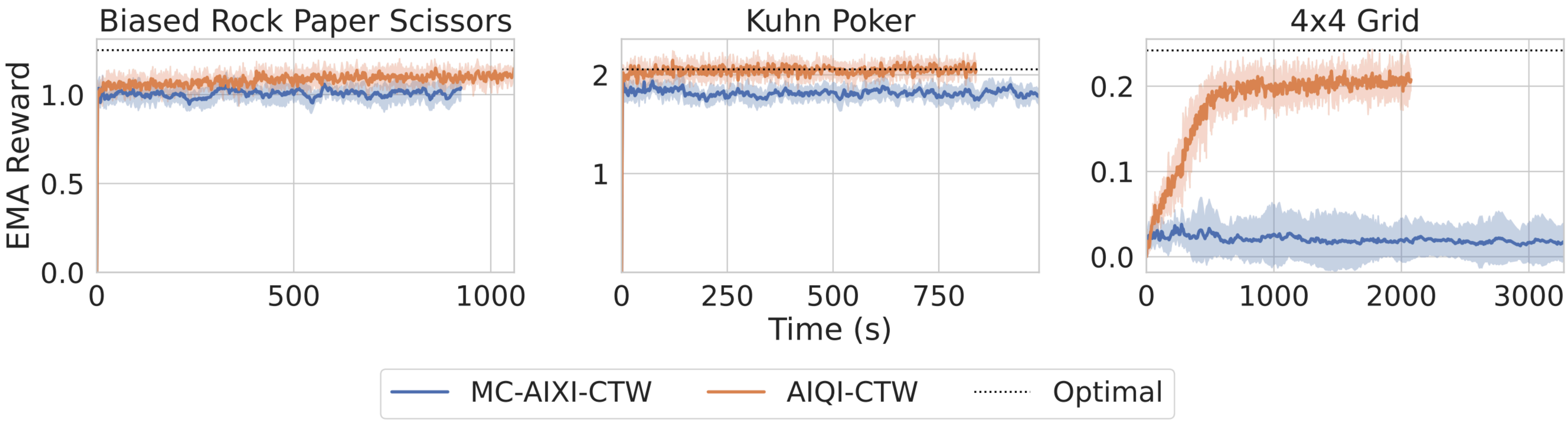
There exist π', $\mathscr{M}$ that admits a self-optimizing policy but AIQI is not even ε-self-optimizing (Theorem 4.10)

Basic idea:



Experimental Results

Under a computational budget, AIQI-CTW outperforms MC-AIXI-CTW



Future Work

- Better exploration, e.g., Thompson sampling
- TD learning (also related to self-optimizing property)
- Extension to semi-measures