AIXI: Universal Optimal Sequential Decision Making

Marcus Hutter (2005)
Reinforcement Learning

• State space $S$, Action space $A$, Policy $\pi$, Reward $R(a, s)$

• Goal: Find policy which maximizes expected cumulative reward.

• Challenge: Environment which RL interacts with is unknown
  • Explore and approximate the environment
  • Hard to balance exploration vs exploitation

• AIXI: why approximate one environment? Consider them all!
Optimal Agents in Known Environments

• \((\mathcal{A}, \mathcal{O}, R) = (\text{action, observation, reward})\) spaces
  • \(a_k = \text{action at time } k\), \(x_k = o_k r_k = \text{perception at time } k\)
• Agent follows policy \(\pi: (\mathcal{A} \times \mathcal{O} \times R)^* \rightarrow \mathcal{A}\)
• Environment reacts with \(\mu: (\mathcal{A} \times \mathcal{O} \times R)^* \times \mathcal{A} \rightarrow \mathcal{O} \times R\)
Agent-Environment Visualization

Agent $\pi$

Environment $\mu$

$r_1 | o_1$, $r_2 | o_2$, $r_3 | o_3$, $r_4 | o_4$, $r_5 | o_5$, $r_6 | o_6$, ...

$a_1$, $a_2$, $a_3$, $a_4$, $a_5$, $a_6$, ...

mem-ory ...

work

tape ...
Optimal Agents in Known Environments

• Performance of $\pi$ is expected cumulative reward

$$V^\pi_\mu = \mathbb{E}^\pi_\mu \left[ \sum_{t=1}^M r_t^{\mu,\pi} \right]$$

• If $\mu$ is true environment, optimal policy is $p^\mu \equiv \arg \max_\pi V^\pi_\mu$
Definition of the Environment

• An environment, \( \rho \), is a sequence of conditional probability functions \( \{ \rho_0, \rho_1, \rho_2, \ldots \} \) and is unknown to the agent.

• Each element in the sequence satisfies the “chronological condition”:

\[
\forall a_{1:n} \forall x_{1:n-1}:
\rho_{n-1}(x_{1:n-1}|a_{1:n-1}) = \sum_{x_n \in X} \rho_n (x_{1:n}|a_{1:n})
\]
Definition of the Environment

\[ \forall a_{1:n} \forall x_{1:n-1} : \rho_{n-1}(x_{1:n-1} | a_{1:n-1}) = \sum_{x_n \in X} \rho_n \binom{x_{1:n} | a_{1:n}} \]

Conditioned on all actions up to \( n - 1 \)

Marginalization of \( \rho_n \) over the current observation-reward

Conditioned on all actions up to \( n \)
Dealing with the Unknown Environment

• The idea is to maintain a *mixture* of environment models, in which each model is assigned a weight that represents the agent’s confidence in what it believes is the true environment.

• As the agent obtains more experience, it updates the weights and thus its belief of the underlying environment.

• Reminiscent of a Bayesian agent.
Mixture Model

- $\mathcal{M} \triangleq \{\rho_1, \rho_2, \ldots, \rho_n\}$ is the countable class of environments
- $w_0^\rho > 0$ is the weight assigned to each $\rho \in \mathcal{M}$ such that $\sum_{\rho \in \mathcal{M}} w_0^\rho = 1$

$$
\xi(x_{1:n} | a_{1:n}) \triangleq \sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{1:n} | a_{1:n})
$$
Selecting a Universal Prior

• Occam’s Razor: The simplest solution is the most likely
• Formalized as Kolmogorov Complexity

\[ \xi(x_{1:n}|a_{1:n}) \triangleq \sum_{\rho \in \mathcal{M}} w_0^\rho \rho(x_{1:n}|a_{1:n}) \]
Kolmogorov Complexity

- Length of the shortest program on a Universal Turing Machine which specifies an object
  - In our domain: shortest program which produces environment $\rho$

$$K(\rho) := \min_p \{\text{length}(p) : U(p) = \rho\}$$

- Advantage: completely independent of prior assumptions
- Problem: Incomputable due to halting problem.
  - Naïve search over all inputs will contain those with infinite loops
  - Paradoxical: “Shortest object describable by N bits” is less than N bits.
Solomonoff Prior

• Key idea: Use inverse Kolmogorov Complexity as environmental prior to compute mixture over all possible environments

\[ \Upsilon(\pi) = \sum_{\rho \in M_U} 2^{-K(\rho)} \ast V^\pi_\rho \]

• \( \Upsilon(\pi) \) measures agent’s ability to perform in all possible environments
• Hutter describes this \( \Upsilon(\pi) \) as Universal Intelligence
AIXI

\[ a_t^{AIXI} = \arg \max_{a_t} \sum_{x_t} \ldots \max_{a_{t+m}} \sum_{x_{t+m}} \left[ \sum_{i=t}^{t+m} r_i \right] \sum_{\rho \in \mathcal{M}} 2^{-K(\rho)} \rho(x_{1:t+m} | a_{1:t+m}), \]

- Expectimax over Solomonoff Prior
- \( \mathcal{M} \) are chronologically conditional environments
- Converges to agent acting with knowledge of true environment
  - Mathematically proven
Evaluation: Pros and Cons

- Theoretically optimal decision making.
  - Proven to converge to optimal agent acting in true environment
- Universal
  - Prior completely independent of actual environment behavior
  - “Reduces any conceptual AI problem to computation problem”
- Incomputable and Intractable
  - Cannot compute Kolgomorov Complexity
- Reward function?
  - Unclear how to define reward function which is also independent of problem
Related Works: Approximations

• Work in AIXI mainly in approximating the theoretical framework.

• AIXItl
  • Summary: provides approximate AIXI which is more optimal than any other RL agent with the same time and space constraints.

• MC-AIXI (Next!)
  • Summary: Monte Carlo approximation of AIXI.
MC-AIXI CTW

• “Monte Carlo – AIXI with Context Tree Weightings”
  • Veness et al 2011

\[
a^\text{AIXI}_t = \arg \max_{a_t} \sum_{x_t} \ldots \max_{a_{t+m}} \sum_{x_{t+m}} \left[ \sum_{i=t}^{t+m} r_i \right] \sum_{\rho \in \mathcal{M}} 2^{-K(\rho)} \rho(x_{1:t+m} | a_{1:t+m})
\]

• Solves main barriers to applying AIXI:
  1. Expectimax is intractable → Estimate using MCTS
  2. Kolmogorov Complexity is incomputable → Replace universe of environments with smaller model class with surrogate for complexity
Part 1: MCTS

• $\rho$UCT is used to estimate AIXI Expectimax by adapting the classic selection-expansion-rollout-backprop MCTS algorithm

• Decision node (circle):
  • Contains a history, $h$, and a value function estimate, $\hat{V}(h)$
  • It has children (called “Chance nodes”) corresponding to the number of possible actions
  • An action, $a$, is selected based on the UCB action-selection policy that balances exploration and exploitation

• Chance node (star):
  • Follows a decision node
  • Contains the history, $ha$; an estimate of the future value, $\hat{V}(ha)$; and the environment model, $\rho(\cdot | ha)$, that returns a percept conditioned on the history
  • A new child of the chance node is added when a new percept is received
Part 2: Approximating the Solomonoff Prior

• Solomonoff Prior: $\sum\rho 2^{-K(\rho)}$ is incomputable

• Solution: Replace with smaller class of environments

• Variable Order Markov Process
  • Calculates probability of next observation depending on last k observations
  • Replace entire universe of environments with mixture of Markov Processes
Prediction Suffix Tree

- Representation of a sequence of binary events
- Able to encode all variable order Markov Models up to depth D
- Represents a space of $2^{2^D}$
Context Tree Weighting

• Provides method to evaluate PST in linear time
  • Naively computable in $\mathcal{O}(2^{2^D})$, CTW algorithm reduces to $\mathcal{O}(D)$

• Smaller trees represent simpler Markov Models
  • Evaluate prior probability under Occam’s razor as size of tree

\[ \Gamma_D(M) = \# \text{ nodes in PST} \]

• Replace Kolmogorov prior with CTW prior
Context Tree Weighting: Updated Formula

• Original intractable prior

\[ a_t^* = \arg \max_{a_t} \sum_{x_t} \ldots \max_{a_{t+m}} \sum_{x_{t+m}} \left[ \sum_{i=t}^{t+m} r_i \right] \sum_{\rho \in \mathcal{M}} 2^{-K(\rho)} \rho(x_{1:t+m} | a_{1:t+m}), \]

• MC-AIXI with CTW

\[ \arg \max_{a_t} \sum_{x_t} \ldots \max_{a_{t+m}} \sum_{x_{t+m}} \left[ \sum_{i=t}^{t+m} r_i \right] \sum_{M \in \mathcal{C}_{D_1} \times \cdots \times \mathcal{C}_{D_k}} 2^{-\sum_{i=1}^{k} \Gamma_{D_i}(M_i)} \Pr(x_{1:t+m} | M, a_{1:t+m}). \]
Algorithm Performance

Cheese Maze

- The agent must navigate to a piece of cheese
- -1 for entering an open cell
- -10 for hitting a wall
- +10 for finding cheese

Partially Observable Pacman

- Agent is unaware of the monsters’ locations and the maze
- It can only “smell” food and observe food in its direct line of sight
Performance on Cheese Maze

Learning Scalability - Cheese Maze

- MC-AIXI
- U-Tree
- Active-LZ
- Optimal

Average Reward per Cycle

Experience (cycles)
Performance on PO-Pacman

[Diagram showing the scaling properties of Partially Observable Pacman with different numbers of simulations (500, 1000, 2000, 5000) over experience cycles.]
Related Work


Timeline

Solomonoff Induction
Ray Solomonoff 1960’s

Context Tree Weightings
Willems, Shtarkov, Tjalkens 1995

MCTS “Bandit based MC Planning”
Kocsis & Szepesvari 2006

MC-AIXI-CTW
Veness et al 2010

Kolmogorov Complexity
Andrey Kolmogorov 1963

AIXI
Marcus Hutter 2005

AIXItl
Marcus Hutter 2007

Kolmogorov Complexity
Andrey Kolmogorov 1963

AIXI
Marcus Hutter 2005

AIXItl
Marcus Hutter 2007
MC-AIXI-CTW Playing Pac-Man

- jveness.info/publications/pacman_jair_2010.wmv