Abstract
Reinforcement Learning agents are expected to eventually perform well. Typically, this takes the form of a guarantee about the asymptotic behavior of an algorithm given some assumptions about the environment. We present an algorithm for a policy whose value approaches the optimal value with probability 1 in all computable probabilistic environments, provided the agent has a bounded horizon. This is known as strong asymptotic optimality, and it was previously unknown whether it was possible for a policy to be strongly asymptotically optimal in the class of all computable probabilistic environments.

Our agent, Inquisitive Reinforcement Learner (Inq), is more likely to explore the more it expects an exploratory action to reduce its uncertainty about which environment it is in, hence the term inquisitive. Exploring inquisitively is a strategy that can be applied generally; for more manageable environment classes, inquisitiveness is tractable. We conducted experiments in "grid-worlds" to compare the Inquisitive Reinforcement Learner to other weakly asymptotically optimal agents.

RL in General Environments
The "environment" is a probability distribution over observation and reward \((\mathcal{O} \times \mathcal{R})\) given all prior actions, observations, and rewards. These probabilities are computable.

Exploratory Expeditions
- Bayesian RL agent with a prior over all computable environments
- Explores as Knowledge Seeking Agent \([\text{OLH13}]\)
- i.e. explores to maximize information gain
- \(m\)-step information gain = how poorly current posterior over environments approximates posterior after \(m\) steps (using KL-divergence)
- \(m\)-k expedition is the \(m\)-step-info-gain-maximizing policy that began \(k\) steps ago

Exploratory Expeditions
- Bayesian RL agent with a prior over all computable environments
- Explores as Knowledge Seeking Agent \([\text{OLH13}]\)
- i.e. explores to maximize information gain
- \(m\)-step information gain = how poorly current posterior over environments approximates posterior after \(m\) steps (using KL-divergence)
- \(m\)-k expedition is the \(m\)-step-info-gain-maximizing policy that began \(k\) steps ago

Central Result
- Our agent’s policy’s value approaches the optimal value in any computable environment.
- No finite-state Markov or ergodicity assumption.

Asymptotic Optimality
Value of a policy: expected future discounted reward (given an interaction history)

Strong Asymptotic Optimality: policy’s value approaches optimal value with probability 1.
Weak Asymptotic Optimality: policy’s value approaches optimal value in Cesàro average with probability 1. \([\text{LH11}]\)

Asymptotic Optimality (Formally)
Value of policy \(\pi\) in environment \(\nu\) after interaction history \(h_{c:t}\):

\[
V^\pi(h_{c:t}) := \frac{1}{\Gamma(t)} \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\left[ r_k | h_{c:t} \right]
\]

Strong Asymptotic Optimality: for all computable environments \(\mu\),

\[
V^\pi(h_{c:t}) - V^\mu(h_{c:t}) \xrightarrow{t \to \infty} 0
\]

with \(\mathbb{P}^\mu\text{-prob. 1}

Inquisitive Reinforcement Learner (Inq)
- Follow the \(m\)-k exploratory expedition with probability proportional to expected info-gain (but capped at \(1/(m!(m+t)))\).
- Else: exploit as a Bayesian reinforcement learner.

Inq is Strongly Asymptotically Optimal
Inq is the first strongly asymptotically optimal agent in general environments, provided it has a "bounded horizon", i.e. doesn’t become more and more farsighted.\(^\text{40}\)

\(\text{40}\) Geometric discounting gives a bounded horizon.

\(\text{40}\) Formally bounded discounting means \(\forall t \exists m \forall \nu : (\sum_{k=m+1}^{\infty} \gamma^k)/(\sum_{k=1}^{m} \gamma^k) \leq \epsilon\)

References

Experiments
Our agent can be made tractable using a smaller model class.

Gridworld environment. Model class is that the reward dispenser could be at any accessible square. Green is agent’s posterior probability reward dispenser there.

Average reward in 10 × 10 (top) and 20 × 20 (bottom) gridworlds. Both baseline agents are weakly asymptotically optimal. Following \([\text{AS17}]\), \(\mu\text{UCT}\) replaces expectimax and the planning horizon is restricted.

Motivating Claim: environments that enter completely novel states are hard to model as MDP:
- chatbot
- function optimizer
- theorem prover

as a conversation with a person progresses, the person never returns to the same state, even approximately.

Inquisitive Reinforcement Learner (Inq)
- Follow the \(m\)-k exploratory expedition with probability proportional to expected info-gain (but capped at \(1/(m!(m+t)))\).
- Else: exploit as a Bayesian reinforcement learner.

Inq is Strongly Asymptotically Optimal
Inq is the first strongly asymptotically optimal agent in general environments, provided it has a "bounded horizon", i.e. doesn’t become more and more farsighted.\(^\text{40}\)