A Strongly Asymptotically Optimal Agent in General Environments

Michael K. Cohen
michael-k-cohen.com

Elliot Catt
ejcatt.github.io

Marcus Hutter
hutter1.net

Australian National University

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.

Exploring When Completely Novel States Abound

Motivating Claim: environments that enter completely novel states infinitely often render (PO)MDP-inspired exploration strategies helpless.

Example environments hard to model as MDP:

- chatbot[a]
- function optimizer
- theorem prover

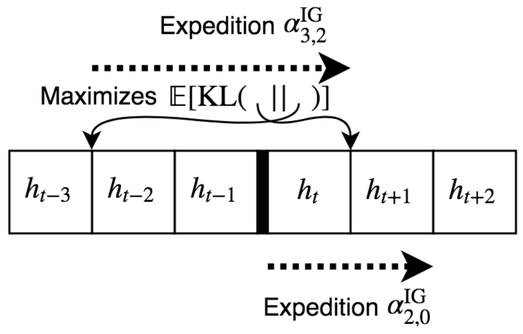
[a] as a conversation with a person progresses, the person never returns to the same state, even approximately.

Central Result

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

Exploratory Expeditions

- Bayesian RL agent with a prior over all computable environments
- Explores as **Knowledge Seeking Agent** [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



Inquisitive Reinforcment Learner (Inq)

- Follow the *m*-*k* exploratory expedition with probability proportional to expected info-gain (but capped at $\frac{1}{m^2(m+1)}$).
- Else: exploit as a Bayesian reinforcement learner.

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 Cesáro average with probability 1. [LH11]

Asymptotic Optimality (Formally)

Value of policy π in environment ν after interaction history $h_{<t}$:

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

Strong Asymptotic Optimality: for all computable environments μ,

$$V^{*}_{\mu}(h_{<t}) - V^{\pi}_{\mu}(h_{<t}) \overset{t\to\infty}{\to} 0 \quad \text{with } \mathrm{P}^{\pi}_{\mu}\text{-prob. } 1$$

Inq is Strongly Asymptotically Optimal

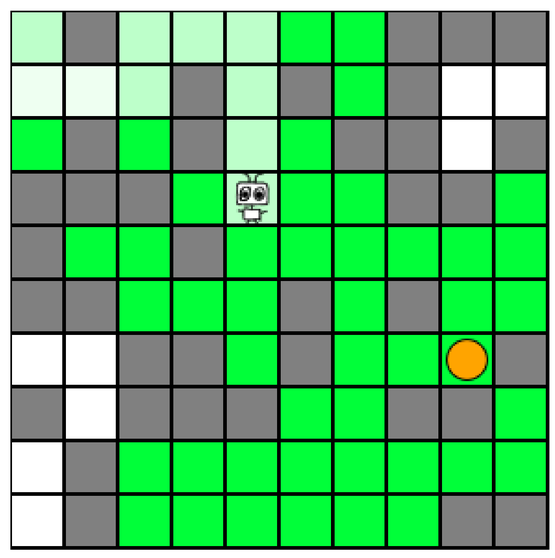
Inq is the first strongly asymptotically optimal agent in general environments, provided it has a "bounded horizon", i.e. doesn't become more and more farsighted.[ab]

[a] Geometric discounting gives a bounded horizon.

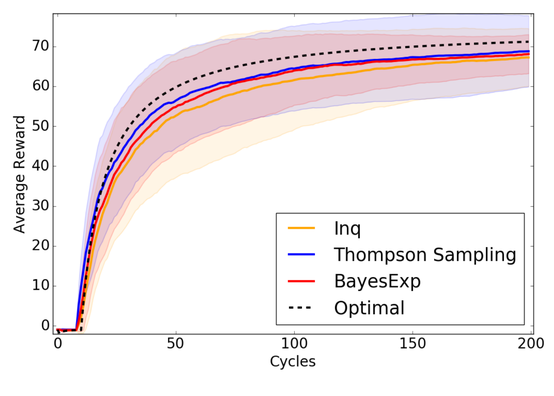
[b] Formally bounded horizon means $\forall\varepsilon\ \exists m\ \forall t : (\sum_{k=t+m}^{\infty}\gamma_k)/(\sum_{k=t}^{\infty}\gamma_k) \le \varepsilon$

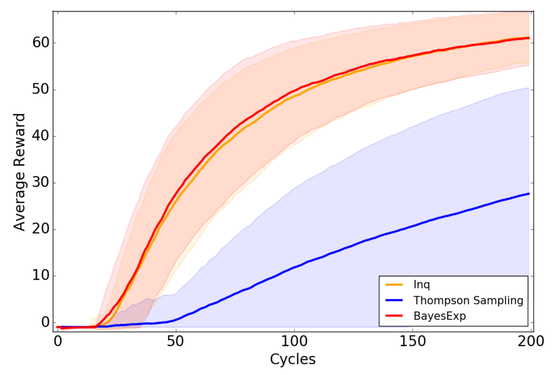
Experiments

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 is there.





Average reward in 10 × 10 (top) and 20 × 20 (bottom) gridworlds. Both baseline agents are weakly asymptotically optimal. Following [Asl17], ρUCT replaces expectimax and the planning horizon is restricted.

References

[Asl17] John Aslanides. AIXIjs: A software demo for general reinforcement learning. *arXiv preprint arXiv:1705.07615*, 2017.

[LH11] Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 368–382, Espoo, Finland, 2011. Springer.

[OLH13] Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In *Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)*, volume 8139 of *LNAI*, pages 158–172, Singapore, 2013. Springer.