A Strongly Asymptotically Optimal Agent in General Environments

Michael K. Cohen, Elliot Catt, Marcus Hutter

Australian National University

16 August 2019

Central result

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

Exploitation vs. exploration

When should you:

* Go to your favorite restaurant
* Fund space travel using current materials
* Sell trinkets where you've had the best luck

* Try a new restaurant
* Fund materials science
* Revisit another place

"Efforts to solve [an instance of the exploration-exploitation problem] so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage." –Peter Whittle

Motivation: exploring when novel states abound

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

Example environments hard to model as an MDP:

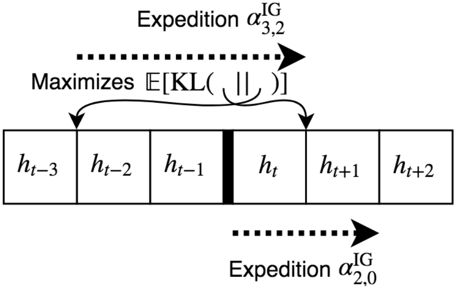
* chatbot
* function optimizer
* theorem prover

Bayesian Reinforcement Learning

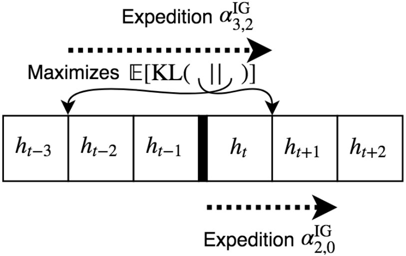
* Start with a prior distribution over what the environment is
* Update this into a posterior distribution
* Maximize expected reward using current "beliefs"

Exploratory Expeditions

- Explore 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 Reinforcement Learner (Inq)



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

Strong asymptotic optimality

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

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

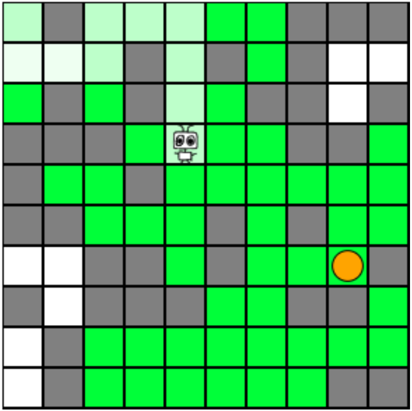
Strong asymptotic optimality: for all computable environments μ,

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

Main Result:
For an agent with a bounded horizon[a], Inq's policy π is strongly asymptotically optimal.

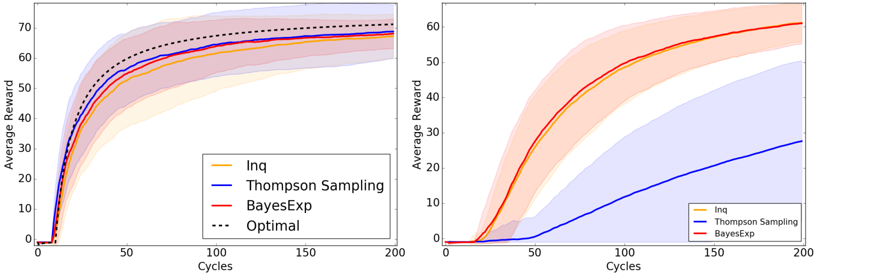
[a] bounded horizon = not becoming more farsighted over time; formally, $\forall \varepsilon \; \exists m \; \forall t : (\sum_{k=t+m}^\infty \gamma_k)/(\sum_{k=t}^\infty \gamma_k) \leq \varepsilon$

Experiments



Gridworld environment. Reward dispensed with probability 3/4 at ●*. Model class is that the reward dispenser could be at any accessible square. Green is agent's posterior probability reward dispenser is there.*

Experimental results



Average reward accumulated in 10x10 (left) and 20x20 (right) gridworlds. Inq is compared to weakly asymptotically optimal agents.

We approximate Inq tractably by replacing expectimax with ρUCT, and restricting the planning horizon.

Thank you