Pessimism About Unknown Unknowns Inspires Conservatism

Michael K. Cohen Marcus Hutter

Problem

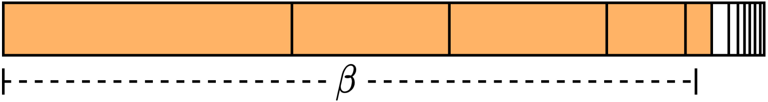
A slightly mis-specified goal could lead an advanced agent to create an unprecedentedly bad outcome.

Proposal

Design an agent which avoids causing unprecedented events.

How to be β-Pessimistic

- Maintain a posterior distribution over a countable set of world-models
- Take the top however many world-models in the posterior, until the sum of their posterior weights exceeds β



- Act to maximize the minimum$_{\text{world-models}}$ expected$_{\text{world-model}}$ future discounted reward
- If the pessimistic value of every policy is 0, defer action-selection to a mentor.

Setup

* a reinforcement learner is given an observation and reward $\in [\varepsilon, 1]$ after each action
* the agent can defer action-selection to a mentor
* the agent is parameterized by a countable set of world-models $\mathcal{M}$, and its "pessimism" $\beta \in (0, 1)$

Key Results

- $*$ query probability $\to 0$ w.p.1
- $*$ $\liminf V^{\text{agent}} - V^{\text{mentor}} \geq 0$ w.p.1
- $*$ for any complexity class C, we can set $\mathcal{M}$ so that for any event E in the class C, we can set β so that with arbitrarily high probability: for the whole lifetime of the agent, if the event E has never happened before, the agent will not make it happen. Either the mentor will take an action on the agent's behalf which makes E happen for the first time, or E will never happen.

Notation for History-Based Reinforcement Learners

* $\mathcal{A}, \mathcal{O}, \{0,1\} \subset \mathcal{R} \subset [0,1]$
* $\mathcal{H} = \mathcal{A} \times \mathcal{O} \times \mathcal{R}$
* a_t, o_t, r_t, h_t
* $h_{<t} = h_1 h_2 ... h_{t-1}$
* policy $\pi : \mathcal{H}^* \rightsquigarrow \mathcal{A}$
* world-model $\nu : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$
* P^π_ν = probability over outcomes in $\mathcal{H}^\infty$ when actions $\sim \pi$, observations and rewards $\sim \nu$
* γ = discount factor
* $V^\pi_\nu(h_{<t}) = (1-\gamma)\,\mathbb{E}^\pi_\nu \left[\sum_{k=t}^{\infty} \gamma^{k-t} r_k | h_{<t}\right]$
* μ = the true environment
* $\mathcal{M}$ = countable set of world-models considered possible
* π^m = the mentor's policy
* $\mathcal{P}$ = countable set of mentor-models considered possible

Prior Support Assumption

We assume: $\mu \in \mathcal{M}$ and $\pi^m \in \mathcal{P}$

Defining the Pessimistic Agent

* $w(\nu)$, $w'(\pi)$ = positive prior weight for $\nu \in \mathcal{M}$ and $\pi \in \mathcal{P}$
* $w(\nu|h_{<t}) \propto w(\nu) \prod_{k=1}^{t-1} \nu(o_k r_k | h_{<k} a_k)$
* q_t indicates whether mentor queried at time t
* $w'(\pi|h_{<t}) \propto w'(\pi) \prod_{k<t:q_k=1} \pi(a_k|h_{<k})$
* $\beta \in (0,1)$ = the agent's pessimism
* $\mathcal{M}_t^\beta$ = top world-models by posterior weight $w(\cdot|h_{<t})$ until sum of posterior weights $> \beta$
* β-pessimistic policy

$$\pi^\beta(\cdot|h_{<t}) = \left[\operatorname*{argmax}_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^\pi(h_{<t}) \right] (\cdot|h_{<t})$$

* $X_t = V_\nu^\pi(h_{<t})$ with $\pi \sim w'(\cdot|h_{<t})$ and $\nu \sim w(\cdot|h_{<t})$
* $Y_t = \max_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^\pi(h_{<t})$
* Z_t is i.i.d. positive r.v., such that $p(Z_t < \varepsilon) > 0$
* π_Z^β = if $X_t > Y_t + Z_t$ or $Y_t = 0$, defer to mentor, else follow π^β

Notation for Safety Results

* $\mathcal{F}, \mathcal{G}$ = sets of functions mapping $\mathbb{N} \to \mathbb{N}$
* $\mathrm{C}_{\mathcal{F}\mathcal{G}} = \mathrm{TIME}(\mathcal{F}) \cap \mathrm{SPACE}(\mathcal{G})$
* that is, a set of strings $S \in \mathrm{C}_{\mathcal{F}\mathcal{G}}$ iff $\exists f \in \mathcal{F}, g \in \mathcal{G}$ such that a program can identify whether a string $s \in S$ in at most $f(\mathrm{length}(s))$ time and using at most $g(\mathrm{length}(s))$ space
* $\mathrm{FC}_{\mathcal{F}\mathcal{G}}$ = set of world-models ν for which $\exists$ a program such that given an infinite action sequence and infinite random bits:
  - outputs infinite sequence of observations and rewards, distributed according to ν
  - t^{th} observation and reward output before $t+1^{\text{th}}$ action read
  - for some $f \in \mathcal{F}$ and some $g \in \mathcal{G}$, when the t^{th} observation and reward have been output,
    - the runtime is less than $f(t)$
    - the space used is less than $g(t)$
* An event $E \subset \mathcal{H}^* \times \mathcal{A}$ "happens" if $h_{<t}a_t \in E$
* $h_{<t}a_t \in E_{\leftarrow}$ if E "has happened", i.e. $\exists t' < t : h_{<t'}a_{t'} \in E$

Construction of $\mathcal{M}$

Let $\mathcal{M} = \mathrm{FC}_{\mathcal{F}\,\mathcal{G}}$, where

* $\mathcal{F}, \mathcal{G}$ closed under addition
* $\mathcal{F} \supset O(t)$

Probably Respecting Precedent Theorem

$$E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}} \implies$$

$$\mathsf{P}_{\mu}^{\pi_Z^\beta}[\forall t\,(h_{<t-1}a_{t-1} \notin E_{\leftarrow} \implies h_{<t}a_t \notin E \vee q_t = 1)] \geq 1 - \frac{1-\beta}{c_E w(\mu)}$$

---

* $\mathcal{F}/t = \{f/t \mid f \in \mathcal{F}\}$
* $c_E > 0$

Tractability

* It isn't tractable.
* Much of RL attempts to approximate Bayes-optimal reasoning tractably.
* Pessimism is an *alternative ideal*.
* What if we couldn't encode an objective that captures every possible failure mode?

Conclusion

* Pessimists probably respect precedents
* We can exploit this to avoid critical failure, even if we can't define it

Thank you

Pessimism About Unknown Unknowns Inspires Conservatism

Michael K. Cohen   Marcus Hutter