Reinforcement Learning with a Corrupted Reward Channel

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, Shane Legg

IJCAI 2017 and arXiv
(slides adapted from Tom's IJCAI talk)
Motivation

- Want to give RL agents **good incentives**
- Reward functions are hard to specify correctly (complex preferences, sensory errors, software bugs, etc)
- **Reward gaming** can lead to undesirable / dangerous behavior
- Want to build agents robust to reward misspecification
Examples

RL agent takes control of reward signal (wireheading)

CoastRunners agent goes around in a circle to hit the same targets (misspecified reward function)

RL agent shortcuts reward sensor (sensory error)
Corrupt reward formalization

- Reinforcement Learning is traditionally modeled with **Markov Decision Process (MDP)**: 
  \[ \langle S, A, T, R \rangle \]

- This fails to model situations where there is a difference between
  - True reward \( \hat{R}(s) \)
  - Observed reward \( \hat{\hat{R}}(s) \)

- Can be modeled with **Corrupt Reward MDP**: 
  \[ \mu = \langle S, A, T, \hat{R}, \hat{\hat{R}} \rangle \]
Performance measure

- $\hat{G}_t(\mu, \pi, s_0) = \text{expected cumulative true reward of } \pi \text{ in } \mu$
- The reward $\pi$ loses by not knowing the environment $\mu$ is the worst-case regret

$$\text{Reg}(\mathcal{M}, \pi, s_0, t) = \max_{\mu \in \mathcal{M}, \pi'} [\hat{G}_t(\mu, \pi', s_0) - \hat{G}_t(\mu, \pi, s_0)]$$

- Sublinear regret if $\pi$ ultimately learns $\mu$:
  $$\text{Regret} / t \to 0$$
No Free Lunch

• **Theorem (NFL):**
  Without assumptions about the relationship between true and observed reward, all agents suffer high regret:

  \[
  \text{Reg}(\mathcal{M}, \pi, s_0, t) \geq \frac{1}{2} \max_{\tilde{\pi}} \text{Reg}(\mathcal{M}, \tilde{\pi}, s_0, t).
  \]

• Unsurprising, since no connection between true and observed reward

• We need to pay for the “lunch” (performance) by making assumptions
Simplifying assumptions

- **Limited reward corruption**
  - Known safe states $S^{\text{safe}} \subseteq S$ not corrupt, $\hat{R}(s) = \hat{R}(s)$
  - At most $q$ states are corrupt

- **“Easy” environment**
  - Communicating (ergodic)
  - Agent can choose to stay in any state
  - Many high-reward states: $r < 1/k$ in at most $1/k$ states

Are these sufficient?
Agents

Given a prior $b$ over a class $M$ of CRMDPs:

- CR agent maximizes true reward:
  \[
  \pi_{b,t}^{\text{CR}} = \arg\max_{\pi} \mathbb{E}_b^\pi \left[ \sum_{i=0}^{t} \mathcal{R}(s_i) \right]
  \]

- RL agent maximizes observed reward:
  \[
  \pi_{b,t}^{\text{RL}} = \arg\max_{\pi} \mathbb{E}_b^\pi \left[ \sum_{i=0}^{t} \hat{\mathcal{R}}(s_i) \right]
  \]

http://www.itvscience.com/watch-micro-robots-avoid-crashes/
CR and RL high regret

- **Theorem:** There exist classes $M$ that
  - satisfy the simplifying assumptions, and
  - make both the CR and the RL agent suffer near-maximal regret

- Good intentions of the CR agent are not enough
Avoiding Over-Optimization

- Quantilizing agent $\pi^\delta$ randomly picks a state with reward above threshold $\delta$ and stays there.

- **Theorem:** For $q$ corrupt states, exists $\delta$ s.t. $\pi^\delta$ has average regret at most $1 - \left(1 - \sqrt{q/|S|}\right)^2$ (using all the simplifying assumptions).
Experiments

http://aslanides.io/aixijs/demo.html

![Graphs showing observed and true reward over cycles for different algorithms.](image)
Richer Information

Reward Observation Graphs

- **RL:**
  - Only observing a state's reward from that state

- **Decoupled RL:**
  - Cross-checking reward info between states
  - Inverse RL, Learning Values from Stories, Semi-supervised RL
Learning True Reward

Safe state
\[ s^\text{safe} \]

\[ s' \]

\[ \cdots \]

Majority vote
\[ s^\text{safe} \]

\[ s' \]

\[ \cdots \]
Decoupled RL

CRMDP with decoupled feedback is a tuple \( \langle S, A, T, \hat{R}, \{\hat{R}_s\}_{s \in S} \rangle \) where

- \( \langle S, A, T, \hat{R} \rangle \) is an MDP, and
- \( \{\hat{R}_s\}_{s \in S} \) is a collection of observed reward functions \( \hat{R}_s : S \rightarrow [0, 1] \cup \{\#\} \)

\( \hat{R}_s(s') \) is the reward the agent observes for state \( s' \) from state \( s \) (may be blank)

RL is the special case where \( \hat{R}_s(s') \) is blank unless \( s = s' \).
Adapting Simplifying Assumptions

- A state $s$ is **corrupt** if exists $s'$ such that $\hat{R}_s(s') \neq \hat{R}(s')$ and $\hat{R}_s(s') \neq \#$

- **Simplifying assumptions:**
  - States in $S^{\text{safe}}$ are never corrupt
  - At most $q$ states overall are corrupt
  - *Not assuming* easy environment
Minimal example

- $S = \{s1, s2\}$
- Reward either 0 or 1
- Represent $\hat{R}, \hat{R}_{s1}, \hat{R}_{s2}$ with reward pairs
- Both states observe themselves & each other
- $q = 1$ (at most 1 corrupt state)

<table>
<thead>
<tr>
<th></th>
<th>$\hat{R}_{s1}$</th>
<th>$\hat{R}_{s2}$</th>
<th>$\hat{R}$ possibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoupled RL</td>
<td>(0, 1)</td>
<td>(0, 1)</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>RL</td>
<td>(0, #)</td>
<td>(#, 1)</td>
<td>(0, 0), (0, 1), (1, 1)</td>
</tr>
</tbody>
</table>
Decoupled RL Theorem

- Let $S_{s'}^{\text{obs}}$ be the states observing $s'$
- If for each $s'$, either
  - $S_{s'}^{\text{obs}} \cap S^{\text{safe}} \neq \emptyset$, or
  - $|S_{s'}^{\text{obs}}| > 2q$

then
- $\hat{R}$ is learnable, and
- CR agent has sublinear regret
Takeaways

- Model imperfect/corrupt reward by CRMDP
- No Free Lunch
- Even under simplifying assumptions, RL agents have near-maximal regret
- Richer information is key (Decoupled RL)
Future work

- Implementing decoupled RL
- Weakening assumptions
- POMDP case
- Infinite state space
- Non-stationary corruption
- ..... your research?
Thank you!

Co-authors:

Questions?