Dead-ends and Secure Exploration in Reinforcement Learning

Mehdi Fatemi
Shikhar Sharma
Harm van Seijen
Samira Ebrahimi Kahu
What is a dead-end?

- A terminal state is called undesired if it prevents achieving maximum return.
- A state $S_d$ is called a dead-end if all the trajectories starting from $S_d$ reach an undesired terminal state with probability 1 in some finite (possibly random) number of steps.

**NOTE:**
- Undesired terminal states are assumed to be signaled when entered.
- NO such assumption can be made for dead-ends.
- Dead-ends may exist far before undesired terminals.
Problem? (why should we care?)

- Just use standard RL algorithms?
  - If the state-space includes many dead-ends and the positive rewards are distant from initial states, then exploration will become a large obstacle.
What do we need?

Security Condition:

A policy $\eta$ is secure if for any $\lambda \in [0,1]$ the following condition holds:

$$\sum_{s' \in \mathcal{S}_D} T(s, a, s') \geq 1 - \lambda \implies \eta(s, a) \leq \lambda$$
A Solution

Make a new MDP (called exploration MDP) similar to the original MDP but with the following:

1. \( r_e = -1 \) if enter an undesired terminal state and \( r_e = 0 \) otherwise.

2. No discount: \( \gamma_e = 1 \)
Theorem

Let $q_e^*$ be the optimal value function of $\mathcal{M}_e$, Let further $\eta$ be any arbitrary policy that satisfies the following:

$$\eta(s, a) \leq 1 + q_e^*(s, a) \quad \forall (s, a) \in S \times A$$

where $q_e^*(s, \cdot) \neq -1$ at least for one action.

Then $\eta$ is secure.
Some Results
Dead-ends and Secure Exploration in Reinforcement Learning

6:30 -- 09:00 PM
Room: Pacific Ballroom

@mefatemi
aka.ms/fatemi

#112