Exploration Conscious Reinforcement Learning Revisited

Lior Shani*  Yonathan Efroni*  Shie Mannor

Technion Institute of Technology
Why?

- To learn a good policy, an RL agent must explore!
- However, it can cause hazardous behavior during training.

I LOVE $\epsilon$-GREEDY
Why?

- To learn a good policy, an RL agent must explore!
- However, it can cause hazardous behavior during training.
Exploration Conscious Reinforcement Learning

- Objective: Find the optimal policy knowing that exploration might occur

- For example: $\epsilon$-greedy exploration ($\alpha = \epsilon$)

$$
\pi_\alpha^* \in \arg\max_{\pi \in \Pi} E^{(1-\alpha)\pi + \alpha\pi_0} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)
$$
Exploration Conscious Reinforcement Learning

- Objective: Find the optimal policy knowing that exploration might occur

  - For example: $\epsilon$-greedy exploration ($\alpha = \epsilon$)

  $$\pi^*_\alpha \in \arg\max_{\pi \in \Pi} \mathbb{E}^{(1-\alpha)\pi_0+\alpha\pi} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

- Solving the Exploration-Conscious problem = Solving an MDP
- We describe a bias-error sensitivity tradeoff in $\alpha$
Exploration Conscious Reinforcement Learning

- Objective: Find the optimal policy knowing that exploration might occur
Fixed Exploration Schemes (e.g. $\epsilon$-greedy)

Choose Greedy Action $a^{greedy}$

- $a^{greedy} \in \arg\max_a Q^\pi (s, a)$
Fixed Exploration Schemes (e.g. $\epsilon$-greedy)

$\bullet \ a^{greedy} \in \arg\max_{a} Q^{\pi_{\alpha}}(s, a)$

$\bullet \ \text{For } \alpha\text{-greedy: } a^{act} \in \begin{cases} a^{greedy} & \text{w.p. } 1 - \alpha \\ \pi_{0} & \text{else} \end{cases}$
Choose Greedy Action $a^{\text{greedy}}$

Draw Exploratory Action $a^{\text{act}}$

Act $a^{\text{act}}$

• $a^{\text{greedy}} \in \arg\max_a Q^{\pi^a}(s, a)$

• For $\alpha$-greedy: $a^{\text{act}} \in \left\{ \begin{array}{ll} a^{\text{greedy}} & \text{w.p. } 1 - \alpha \\ \pi_0 & \text{else} \end{array} \right\}$
Choose Greedy Action $a^{\text{greedy}}$

Draw Exploratory Action $a^{\text{act}}$

Act $a^{\text{act}}$

Receive $r, s'$

- $a^{\text{greedy}} \in \arg\max_a Q^{\pi^a}(s, a)$

- For $\alpha$-greedy: $a^{\text{act}} \in \begin{cases} a^{\text{greedy}} & \text{w.p. } 1 - \alpha \\ \pi_0 & \text{else} \end{cases}$
Fixed Exploration Schemes (e.g. $\epsilon$-greedy)

- Normally used information: $(s, a^{act}, r, s')$
Fixed Exploration Schemes (e.g. $\epsilon$-greedy)

- Normally used information: $(s, a^{act}, r, s')$
- Using information about the exploration process: $(s, a^{greedy}, a^{act}, r, s')$
Two Approaches – Expected approach

1. Update $Q^{\pi}(s_t, a^\text{act}_t)$

2. Expect that the agent might explore in the next state

$$Q^{\pi}(s_t, a^\text{act}_t) \leftarrow \eta \left( r_t + \gamma \mathbb{E}(1-\alpha)^{\pi+\alpha\pi_0} Q^{\pi}(s_{t+1}, a) - Q^{\pi}(s_t, a^\text{act}_t) \right)$$
Two Approaches – Expected approach

1. Update $Q^\pi(s_t, a^\text{act}_t)$

2. Expect that the agent might explore in the next state

$$Q^\pi(s_t, a^\text{act}_t) += \eta \left( r_t + \gamma \mathbb{E}(1-\alpha)\pi + \alpha \pi_0 Q^\pi(s_{t+1}, a) - Q^\pi(s_t, a^\text{act}_t) \right)$$

- Calculating expectations can be hard.
- Requires sampling in the continuous case!
Two Approaches – Surrogate approach

- Exploration is incorporated into the environment!

1. Update $Q^{\pi}(s_t, a_t^{\text{greedy}})$

2. The rewards and next state $r_t, s_{t+1}$ are given by the acted action $a_t^{\text{act}}$

$$Q^{\pi}(s_t, a_t^{\text{greedy}}) + = \eta \left( r_t + \gamma Q^{\pi}(s_{t+1}, a_{t+1}^{\text{greedy}}) - Q^{\pi}(s_t, a_t^{\text{greedy}}) \right)$$
Two Approaches – Surrogate approach

- Exploration is incorporated into the environment!

1. Update $Q^{\pi^a}(s_t, a_t^{\text{greedy}})$

2. The rewards and next state $r_t, s_{t+1}$ are given by the acted action $a_t^{\text{act}}$

$$Q^{\pi^a}(s_t, a_t^{\text{greedy}}) \leftarrow \eta \left( r_t + \gamma Q^{\pi^a}(s_{t+1}, a_{t+1}^{\text{greedy}}) - Q^{\pi^a}(s_t, a_t^{\text{greedy}}) \right)$$

- NO NEED TO SAMPLE!
Deep RL
Experimental Results

Training

Evaluation

[Graphs showing reward over steps for training and evaluation phases, with different lines representing DDPG, Expected $\sigma$-DDPG, and Surrogate $\sigma$-DDPG.

---

Shani, Efroni & Mannor
Exploration Conscious Reinforcement Learning revisited

12-Jun-19
17/19
Summary

• We define Exploration Conscious RL and analyze its properties.
• Exploration Conscious RL can improve performance over both the training and evaluation regimes.

• Conclusion: Exploration-Conscious RL and specifically, the Surrogate approach, can easily help to improve variety of RL algorithms.
Summary

• We define Exploration Conscious RL and analyze its properties.
• Exploration Conscious RL can improve performance over both the training and evaluation regimes.

• Conclusion: Exploration-Conscious RL and specifically, the Surrogate approach, can easily help to improve variety of RL algorithms.

SEE YOU AT POSTER

#90