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.**

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 : ϵ-greedy exploration ($\alpha = \epsilon$)**

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

Exploration Conscious Reinforcement Learning

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

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

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

- **Solving the Exploration-Conscious problem = Solving an MDP**
- **We describe a bias-error sensitivity tradeoff in α**

Exploration Conscious Reinforcement Learning

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

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy}

- $a^{greedy} \in \mathrm{argmax}_a Q^{\pi^\alpha}(s,a)$

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy} → **Draw Exploratory Action a^{act}**

- $a^{greedy} \in \mathrm{argmax}_a Q^{\pi^\alpha}(s, a)$
- **For α-greedy:** $a^{act} \in \begin{Bmatrix} a^{greedy} & \text{w.p. } 1-\alpha \\ \pi_0 & \text{else} \end{Bmatrix}$

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy} → **Draw Exploratory Action** a^{act} → **Act** a^{act}

- $a^{greedy} \in \mathrm{argmax}_a Q^{\pi^{\alpha}}(s,a)$
- **For α-greedy:** $a^{act} \in \begin{Bmatrix} a^{greedy} & \text{w.p. } 1-\alpha \\ \pi_0 & \text{else} \end{Bmatrix}$

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy} → Draw Exploratory Action a^{act} → Act a^{act} → Recieve r, s'

- $a^{greedy} \in \mathrm{argmax}_a Q^{\pi^\alpha}(s, a)$
- **For α-greedy:** $a^{act} \in \begin{Bmatrix} a^{greedy} & \text{w.p. } 1-\alpha \\ \pi_0 & \text{else} \end{Bmatrix}$

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy} → **Draw Exploratory Action** a^{act} → **Act** a^{act} → **Recieve** r, s'

- **Normally used information:** (s, a^{act}, r, s')

Fixed Exploration Schemes (e.g. ϵ-greedy)

Choose Greedy Action a^{greedy} → **Draw Exploratory Action** a^{act} → **Act** a^{act} → **Recieve** r, s'

- **Normally used information:** $\cancel{(s, a^{act}, r, s')}$
- **Using information about the exploration process:** $\left(s, \color{red}{a^{greedy}}, a^{act}, r, s'\right)$

Two Approaches – Expected approach

1. **Update $Q^{\pi^\alpha}(s_t, a_t^{act})$**
2. **Expect that the agent might explore in the next state**

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

Two Approaches – Expected approach

1. **Update $Q^{\pi^\alpha}(s_t, a_t^{act})$**
2. **Expect that the agent might explore in the next state**

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

- Calculating expectations can be hard.
    - **Requires sampling in the continuous case!**

Two Approaches – Surrogate approach

- **Exploration is incorporated into the environment!**

1. **Update** $Q^{\pi^{\alpha}}\left(s_t, a_t^{greedy}\right)$
2. **The rewards and next state** r_t **,** s_{t+1} **are given by the acted action** a_t^{act}

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

Two Approaches – Surrogate approach

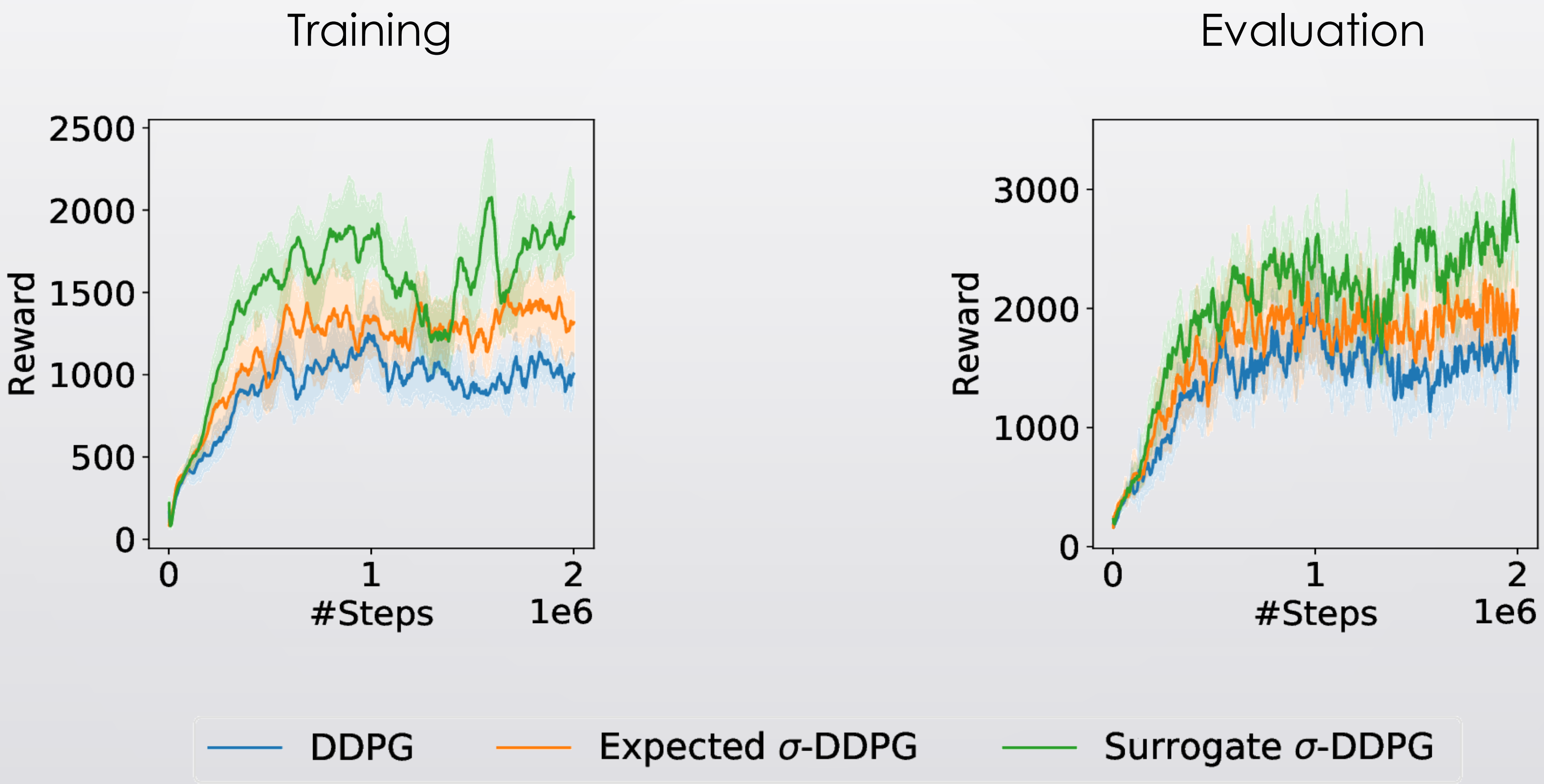
- **Exploration is incorporated into the environment!**

1. **Update** $Q^{\pi^\alpha}\left(s_t, a_t^{greedy}\right)$
2. **The rewards and next state** r_t ,s_{t+1} **are given by the acted action** a_t^{act}

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

- **NO NEED TO SAMPLE!**

Deep RL
Experimental Results



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 [illegible] algorithms.**

SEE YOU AT POSTER

#90