Best of Both Worlds Policy Optimization

Christoph Dann (Google Research), **Chen-Yu Wei (MIT)**, Julian Zimmert (Google Research)

Setting

Setting

- RL in fixed-horizon tabular MDPs
- Reward function is potentially **adversarial**
  - In episode t, the reward function $r_t(s, a)$ is chosen by an adversary
- Regret minimization

Setting

- RL in fixed-horizon tabular MDPs

- Reward function is potentially **adversarial**
  - In episode t, the reward function $r_t(s,a)$ is chosen by an adversary

- Regret minimization

$$\text{Regret} = \text{Total reward of policy } \pi^\star \text{ in } T \text{ episodes} - \text{Total reward of the learner in } T \text{ episodes}$$

Setting

- RL in fixed-horizon tabular MDPs

- Reward function is potentially **adversarial**
  - In episode t, the reward function $r_t(s, a)$ is chosen by an adversary

- Regret minimization

the best fixed policy

$\mathrm{Regret} =$ Total reward of policy $\pi^\star$ in T episodes
– Total reward of the learner in T episodes

Comparison

	Fixed-reward MDP	Adversarial MDP
Value Iteration (Azar et al., 2017, Simchowitz & Jamieson, 2019)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘
Q-learning (Jin et al., 2018, Yang et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘

$$\Delta \coloneqq \min_{s,a} V^\star(s) - Q^\star(s,a)$$

Comparison

	Fixed-reward MDP	Adversarial MDP
Value Iteration (Azar et al., 2017, Simchowitz & Jamieson, 2019)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘
Q-learning (Jin et al., 2018, Yang et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘
Mirror Descent over Occupancy Measure (Jin et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log^2 T}{\Delta}\right\}$	$\sqrt{T}$

$$\Delta := \min_{s,a} V^\star(s) - Q^\star(s,a)$$

Comparison

	Fixed-reward MDP	Adversarial MDP
Value Iteration (Azar et al., 2017, Simchowitz & Jamieson, 2019)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘
Q-learning (Jin et al., 2018, Yang et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✘
Mirror Descent over Occupancy Measure (Jin et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log^2 T}{\Delta}\right\}$	$\sqrt{T}$
Policy Optimization (Luo et al, 2021)	$\sqrt{T}$	$\sqrt{T}$

$$\Delta \coloneqq \min_{s,a} V^\star(s) - Q^\star(s,a)$$

Comparison

	Fixed-reward MDP	Adversarial MDP
Value Iteration (Azar et al., 2017, Simchowitz & Jamieson, 2019)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✗
Q-learning (Jin et al., 2018, Yang et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log T}{\Delta}\right\}$	✗
Mirror Descent over Occupancy Measure (Jin et al., 2021)	$\min\left\{\sqrt{T}, \frac{\log^2 T}{\Delta}\right\}$	$\sqrt{T}$
Policy Optimization (Luo et al, 2021)	$\sqrt{T} \longrightarrow \min\left\{\sqrt{T}, \frac{\log^2 T}{\Delta}\right\}$	$\sqrt{T}$

$$\Delta \coloneqq \min_{s,a} V^\star(s) - Q^\star(s,a)$$

Standard Policy Optimization (e.g., PPO)

$$\pi_{t+1}(\cdot|s) = \max_{\pi} \left\{ \sum_{a} \pi(a|s) \underbrace{\hat{Q}_t(s,a)}_{\text{Q-function estimator for } \pi_t} - \beta D_{\mathrm{KL}}(\pi(\cdot|s), \pi_t(\cdot|s)) \right\}$$

Standard Policy Optimization (e.g., PPO)

$$\pi_{t+1}(\cdot\,|s) = \max_{\pi} \left\{ \sum_{a} \pi(a|s) \underbrace{\hat{Q}_t(s,a)}_{\text{Q-function estimator for } \pi_t} - \beta D_{\mathrm{KL}}(\pi(\cdot\,|s), \pi_t(\cdot\,|s)) \right\}$$

Issue: lack exploration

Standard Policy Optimization (e.g., PPO)

$$\pi_{t+1}(\cdot|s) = \max_{\pi} \left\{ \sum_{a} \pi(a|s) \underbrace{\hat{Q}_t(s,a)}_{\text{Q-function estimator for } \pi_t} - \beta D_{\mathrm{KL}}(\pi(\cdot|s), \pi_t(\cdot|s)) \right\}$$

Issue: lack exploration

The regret scales with **distribution mismatch coefficient** $\max_{t,s} \frac{d^{\pi^\star}(s)}{d^{\pi_t}(s)}$

Policy Optimization with Exploration Bonus

$$\pi_{t+1}(\cdot|s) = \max_{\pi} \left\{ \sum_{a} \pi(a|s) \left(\hat{Q}_t(s,a) + \underbrace{B_t(s,a)}_{\text{exploration bonus}} \right) - \beta D_{\text{KL}}(\pi(\cdot|s), \pi_t(\cdot|s)) \right\}$$

(Luo et al. 2021) Policy optimization in adversarial MDPs: improved exploration via dilated bonuses

Policy Optimization with Exploration Bonus

$$\pi_{t+1}(\cdot|s) = \max_{\pi}\left\{\sum_a \pi(a|s)\left(\hat{Q}_t(s,a) + \underbrace{B_t(s,a)}_{\text{exploration bonus}}\right) - \beta D_{\mathrm{KL}}(\pi(\cdot|s), \pi_t(\cdot|s))\right\}$$

$$B_t(s,a) \approx \mathbb{E}\left[\sum_h \frac{1}{\beta d_t(s_h) + \gamma} \;\middle|\; (s_1, a_1) = (s,a), \quad (s_2, a_2, s_3, a_3 \ldots) \sim \pi_t\right]$$

$$= Q^{\pi_t}\left(s, a;\ \text{reward} = \frac{1}{\beta d_t(s) + \gamma}\right)$$

(Luo et al. 2021)

$d_t(s) \coloneqq$ occupancy measure on state s under policy π_t

(Luo et al. 2021) Policy optimization in adversarial MDPs: improved exploration via dilated bonuses

Policy Optimization with Exploration Bonus

$$\pi_{t+1}(\cdot\,|s) = \max_{\pi}\left\{\sum_a \pi(a|s)\left(\hat{Q}_t(s,a) + \underbrace{B_t(s,a)}_{\text{exploration bonus}}\right) - \beta D_{\mathrm{KL}}(\pi(\cdot\,|s), \pi_t(\cdot\,|s))\right\}$$

$$B_t(s,a) \approx \mathbb{E}\left[\sum_h \frac{1}{\beta d_t(s_h) + \gamma} \,\middle|\, (s_1,a_1) = (s,a), \quad (s_2,a_2,s_3,a_3 \ldots) \sim \pi_t\right]$$

$$= Q^{\pi_t}\left(s,a;\ \text{reward} = \frac{1}{\beta d_t(s) + \gamma}\right)$$

(Luo et al. 2021)

$d_t(s) :=$ occupancy measure on state s under policy π_t

$\Rightarrow$ **No longer suffer from distribution mismatch!**

(Luo et al. 2021) Policy optimization in adversarial MDPs: improved exploration via dilated bonuses

Policy Optimization with Exploration Bonus

(Luo et al. 2021)

Policy optimization with exploration bonus achieves a regret bound of $O\left(H^2 S\sqrt{AT}\right)$ even when the reward is adversarial.

Policy Optimization with Exploration Bonus

(Luo et al. 2021)

Policy optimization with exploration bonus achieves a regret bound of $O\left(H^2 S\sqrt{AT}\right)$ even when the reward is adversarial.

Issue: The bonus leads to over-exploration in fixed-reward MDPs.

Policy Optimization with Exploration Bonus

(Luo et al. 2021)

Policy optimization with exploration bonus achieves a regret bound of $O\left(H^2 S\sqrt{AT}\right)$ even when the reward is adversarial.

Issue: The bonus leads to over-exploration in fixed-reward MDPs.

Sample efficiency in fixed environment ⟷ ? ⟷ Robustness against adversary

Joint Bonus and Regularization Design

$$\pi_{t+1}(\cdot\,|s) = \max_{\pi}\left\{\sum_{a} \pi(a|s)\left(\hat{Q}_t(s,a) + B_t(s,a)\right) - \beta_t(s)\boldsymbol{D}(\pi(\cdot\,|s), \pi_t(\cdot\,|s))\right\}$$

Joint Bonus and Regularization Design

$$\pi_{t+1}(\cdot|s) = \max_{\pi}\left\{\sum_{a}\pi(a|s)\left(\hat{Q}_t(s,a) + B_t(s,a)\right) - \beta_t(s)\boldsymbol{D}(\pi(\cdot|s), \pi_t(\cdot|s))\right\}$$

$\boldsymbol{D} = \boldsymbol{D}_{\mathbf{KL}}$:

$$\beta_{t+1}(s) \leftarrow \beta_t(s) + \frac{\frac{1}{d_t(s)}}{\sqrt{\sum_{\tau=1}^{t}\frac{\psi_\tau(s)}{d_\tau(s)}}} \qquad \psi_t(s) = \mathrm{Entropy}(\pi_t(\cdot|s))$$

$$B_t(s,a) = Q^{\pi_t}\left(s,a;\ \mathrm{reward} = \left(\beta_{t+1}(s) - \beta_t(s)\right)\psi_t(s)\right)$$

Joint Bonus and Regularization Design

$$\pi_{t+1}(\cdot\,|s) = \max_{\pi}\left\{\sum_{a} \pi(a|s)\left(\hat{Q}_t(s,a) + B_t(s,a)\right) - \beta_t(s)\boldsymbol{D}(\pi(\cdot\,|s), \pi_t(\cdot\,|s))\right\}$$

$\boldsymbol{D} = \boldsymbol{D}_{\mathbf{TS}}$ (Bregman divergence defined by ½-Tsallis entropy)**:**

$$\beta_t(s) = \sqrt{\sum_{\tau=1}^{t} \frac{1}{d_\tau(s)}} \qquad \psi_t(s) = \sum_{a} \sqrt{\pi_t(a|s)}\,(1 - \pi_t(a|s))$$

$$B_t(s,a) = Q^{\pi_t}\left(s,a;\ \text{reward} = \left(\beta_{t+1}(s) - \beta_t(s)\right)\psi_t(s)\right)$$

Summary

- In tabular MDPs, policy optimization achieves the **best of both worlds**:
  - Similar to VI or Q-learning in fixed-reward MDPs, but additionally handles adversarial MDPs.
- The key is to jointly design the **exploration bonus** and **regularization term** in an adaptive way.

Summary

- In tabular MDPs, policy optimization achieves the **best of both worlds**:
  - Similar to VI or Q-learning in fixed-reward MDPs, but additionally handles adversarial MDPs.
- The key is to jointly design the **exploration bonus** and **regularization term** in an adaptive way.

Poster: 11am, July 26 (Wed), #622