Provably Efficient Safe Exploration via Primal-Dual Policy Optimization

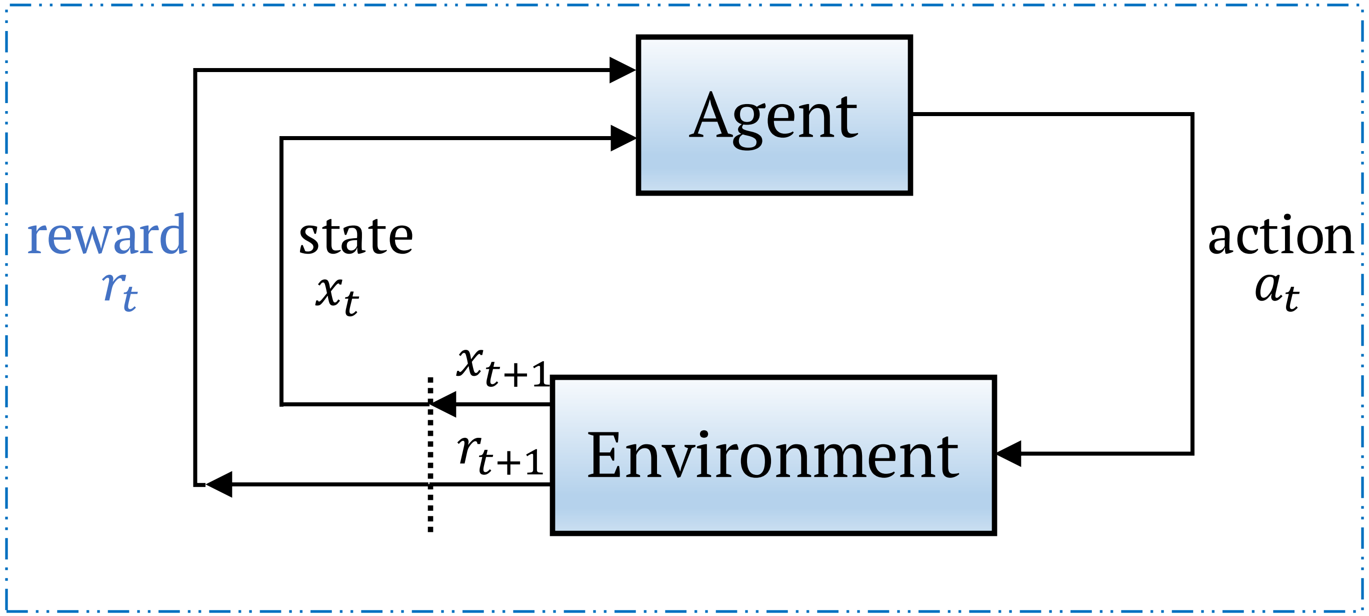
Dongsheng Ding (USC)

a joint work with

Xiaohan Wei (FB)
Zhuoran Yang (Princeton)
Zhaoran Wang (NU)
Mihailo R. Jovanović (USC)

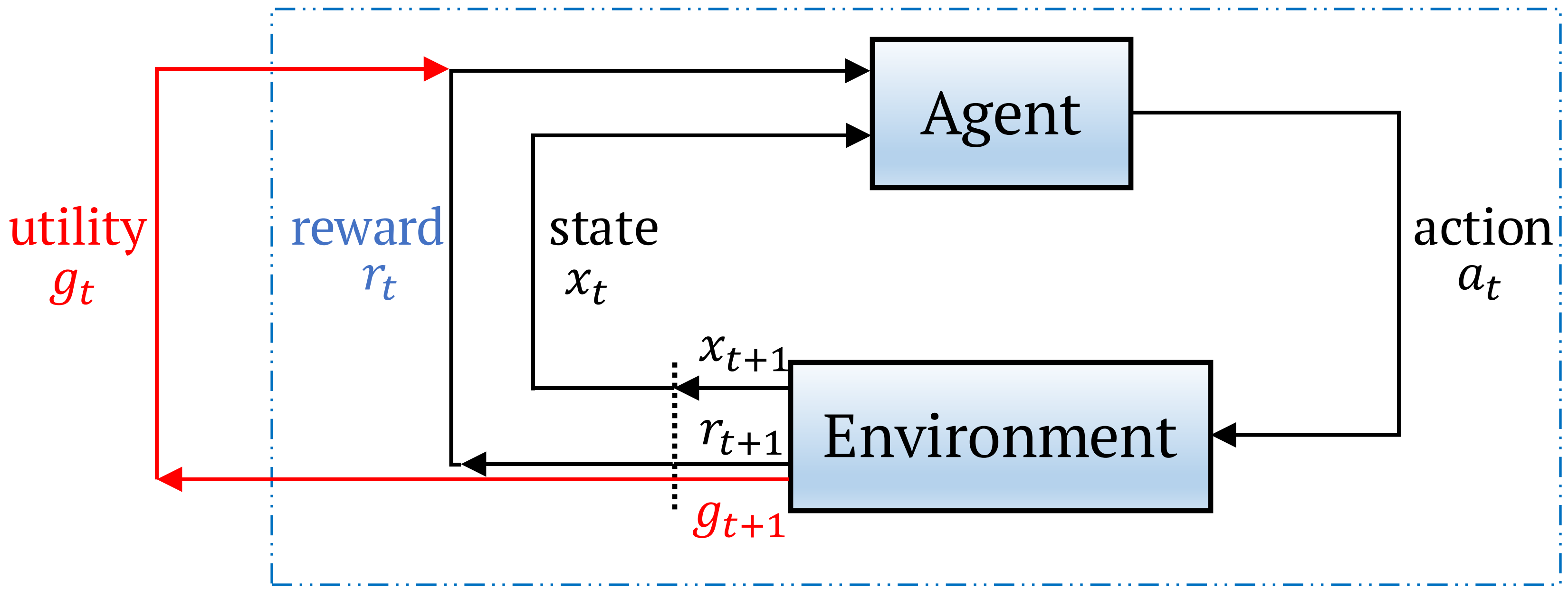
The 24th International Conference on Artificial Intelligence and Statistics, April 13rd -15th , 2021

Constrained Sequential Decision-Making



- Framework: Reinforcement Learning

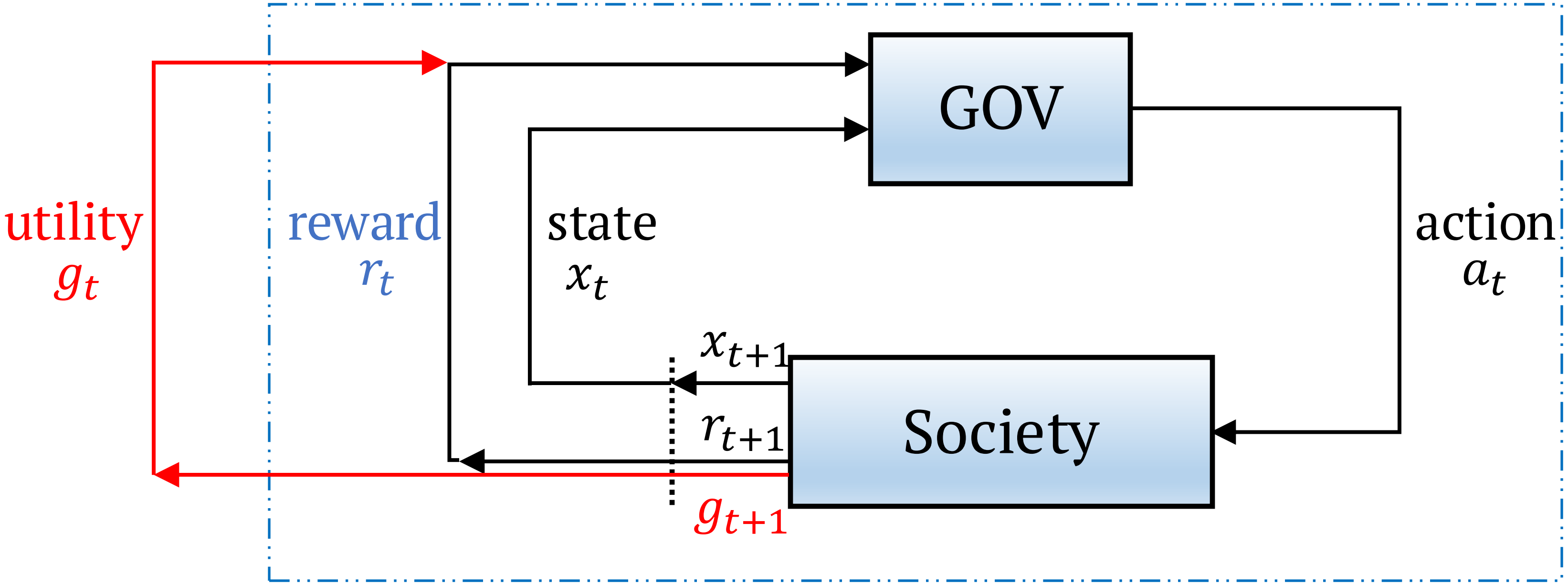
Constrained Sequential Decision-Making



- Framework: Reinforcement Learning
- Add constraints on the utility

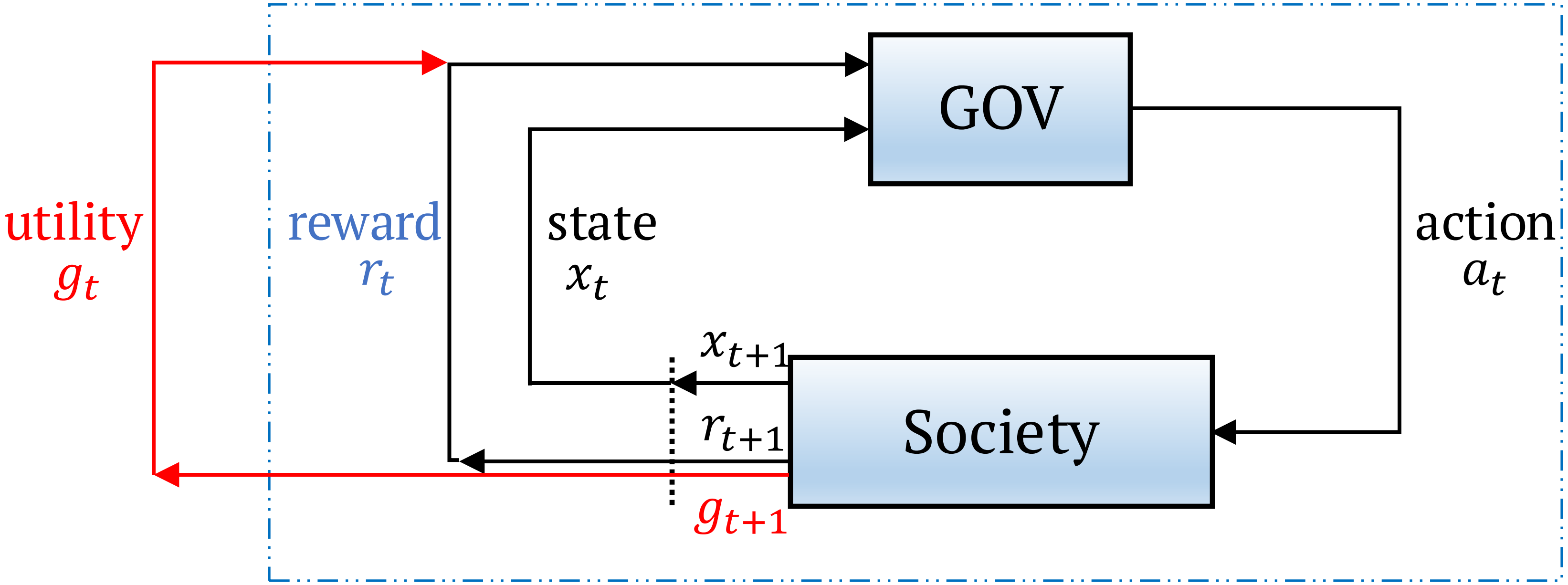
Example: Pandemic Control

Figure: IHRB '20



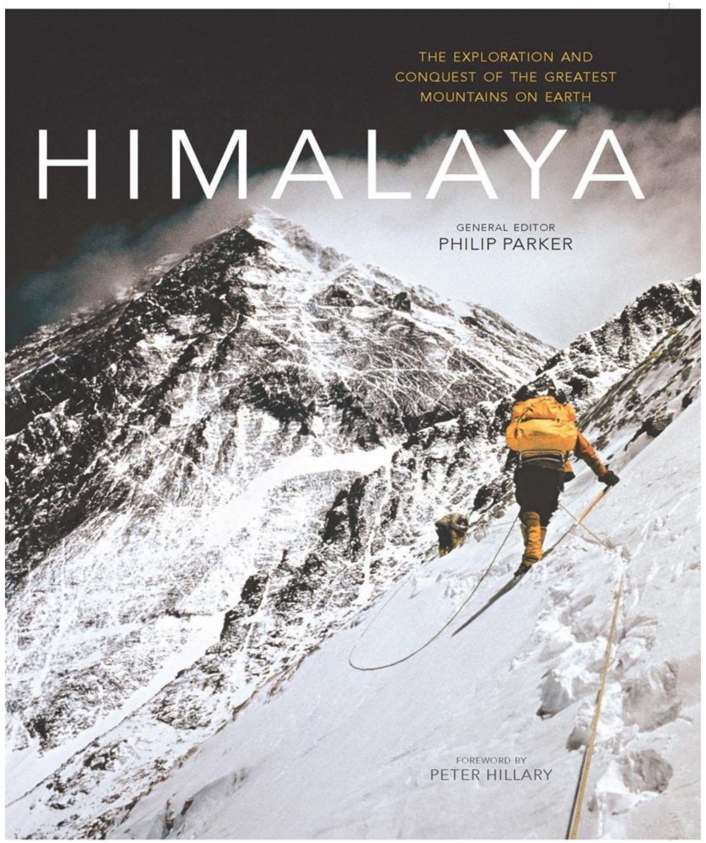
Example: Pandemic Control

Figure: IHRB '20



- In RL, the agent needs to explore the unknown environment.
- Exploration is costly.

Safe Exploration



- Objective #1: maximize the long-term reward.
- Objective #2: maintain the long-term constraint satisfaction.

Environment Model

- (episodic) Constrained MDP / CMDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g)$

$$x_1, \ldots, x_h, a_h \sim \pi_h(\cdot \mid x_h),\ r_h(x_h, a_h),\ g_h(x_h, a_h),\ x_{h+1} \sim \mathbb{P}_h(\cdot \mid x_h, a_h)$$

Altman, '99

Environment Model

- (episodic) Constrained MDP / CMDP ($\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g$)

$$x_1, \ldots, x_h, a_h \sim \pi_h(\cdot \mid x_h),\ r_h(x_h, a_h),\ g_h(x_h, a_h),\ x_{h+1} \sim \mathbb{P}_h(\cdot \mid x_h, a_h)$$

- Find an optimal policy $\pi^\star$ that solves,

$$\underset{\pi}{\text{maximize}} \quad V_{r,1}^{\pi}(x_1)$$

$$\text{subject to} \quad V_{g,1}^{\pi}(x_1) \geq b$$

- $V_{r,1}^{\pi}(x_1) = \mathrm{E}_{\pi}\left[\sum_{h=1}^{H} r_h(x_h, a_h) \mid x_1\right]$
- $V_{g,1}^{\pi}(x_1) = \mathrm{E}_{\pi}\left[\sum_{h=1}^{H} g_h(x_h, a_h) \mid x_1\right]$

Altman, '99

This Work

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting ?

This Work

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting ?

- Online episodic constrained MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g)$

$$\pi^k = \{ \pi_h^k(\cdot \mid \cdot) \}_{h=1}^H, \qquad k = 1, 2, \ldots, K$$

This Work

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting ?

- Provably sample efficient

$$\text{Regret}(K) = \sum_{k=1}^{K} \left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi^k}(x_1) \right) \qquad \text{Violation}(K) = \sum_{k=1}^{K} \left(b - V_{g,1}^{\pi^k}(x_1) \right)$$

This Work

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting ?

Linear Function Approximation

- Kernel feature map $\psi: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^{d_1}$

$$\mathbb{P}_h(x' \mid x, a) = \langle \psi(x, a, x'), \theta_h \rangle$$

- Reward/utility feature map $\varphi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$

$$r_h(x, a) = \langle \varphi(x, a), \theta_{r,h} \rangle \quad \text{and} \quad g_h(x, a) = \langle \varphi(x, a), \theta_{g,h} \rangle$$

- Special cases: finite CMDPs, linear mixture kernel, etc.

Yang, Wang, '20
Ayoub, Jia, Szepesvari, Wang, & Yang, '20
Zhou, He, Gu, '20

This Work

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting ?

Lagrangian-Based Policy Optimization

- Saddle-point problem

$$\underset{\pi}{\text{maximize}} \ \underset{Y \geq 0}{\text{minimize}} \quad \mathcal{L}(\pi, Y) \quad := \quad \underbrace{V_{r,1}^{\pi}(x_1)}_{\text{Objective}} - \underbrace{Y\left(b - V_{g,1}^{\pi}(x_1)\right)}_{\text{Penalty}}$$

- Primal-dual update

$$\pi^k \quad \longleftarrow \quad \text{Gradient Ascent}\left(\pi^{k-1}, Y^{k-1}, \nabla_\pi \mathcal{L}(\pi^{k-1}, Y^{k-1})\right)$$

$$Y^k \quad \longleftarrow \quad \text{Gradient Descent}\left(\pi^{k-1}, Y^{k-1}, \nabla_Y \mathcal{L}(\pi^{k-1}, Y^{k-1})\right)$$

Lagrangian-Based Policy Optimization

- Saddle-point problem

$$\underset{\pi}{\text{maximize}} \ \underset{Y \geq 0}{\text{minimize}} \quad \mathcal{L}(\pi, Y) \quad := \quad \underbrace{V_{r,1}^{\pi}(x_1)}_{\text{Objective}} - \underbrace{Y\left(b - V_{g,1}^{\pi}(x_1)\right)}_{\text{Penalty}}$$

- Primal-dual update

$$\pi^k \quad \longleftarrow \quad \text{Gradient Ascent}\left(\pi^{k-1}, Y^{k-1}, \nabla_\pi \mathcal{L}(\pi^{k-1}, Y^{k-1})\right)$$

$$Y^k \quad \longleftarrow \quad \text{Gradient Descent}\left(\pi^{k-1}, Y^{k-1}, \nabla_Y \mathcal{L}(\pi^{k-1}, Y^{k-1})\right)$$

- Used in AC (Borkar, et al., ’05), RCPO (Tessler, et al., ’19), dualDescent (Paternain, et al., ’19), NPG-PD (Ding, et al., ’20), et al.

Approximate Lagrangian

$$\mathcal{L}(\pi, Y^{k-1}) \quad \approx$$

- Local approximation in TRPO/PPO

$$V_{r,1}^{\pi}(x_1) \quad \approx \quad V_{r,1}^{\pi^{k-1}}(x_1) \;+ \sum_{h=1}^{H} \left\langle Q_{r,1}^{\pi^{k-1}}(x_h,\cdot), \left(\pi_h - \pi_h^{k-1}\right)(\cdot\,|x_h) \right\rangle$$

$$V_{g,1}^{\pi}(x_1) \quad \approx \quad V_{g,1}^{\pi^{k-1}}(x_1) \;+ \sum_{h=1}^{H} \left\langle Q_{g,1}^{\pi^{k-1}}(x_h,\cdot), \left(\pi_h - \pi_h^{k-1}\right)(\cdot\,|x_h) \right\rangle$$

Schulman, et al., ’15 & ’17

Approximate Lagrangian

$$\mathcal{L}(\pi, Y^{k-1}) \approx V_{r,1}^{\pi^{k-1}}(x_1) - Y^{k-1}\left(b - V_{g,1}^{\pi^{k-1}}(x_1)\right) + \sum_{h=1}^{H}\left\langle \left(Q_{r,h}^{\pi^{k-1}} + Y^{k-1} Q_{g,h}^{\pi^{k-1}}\right)(x_h,\cdot), (\pi_h - \pi_h^{k-1})(\cdot\,|x_h) \right\rangle$$

- Local approximation in TRPO/PPO

$$V_{r,1}^{\pi}(x_1) \approx V_{r,1}^{\pi^{k-1}}(x_1) + \sum_{h=1}^{H}\left\langle Q_{r,1}^{\pi^{k-1}}(x_h,\cdot), (\pi_h - \pi_h^{k-1})(\cdot\,|x_h) \right\rangle$$

$$V_{g,1}^{\pi}(x_1) \approx V_{g,1}^{\pi^{k-1}}(x_1) + \sum_{h=1}^{H}\left\langle Q_{g,1}^{\pi^{k-1}}(x_h,\cdot), (\pi_h - \pi_h^{k-1})(\cdot\,|x_h) \right\rangle$$

Schulman, et al., ’15 & ’17

Primal-Dual Proximal Policy Optimization

- Primal update

$$\pi^k \leftarrow \underset{\pi}{\operatorname{argmax}} \sum_{h=1}^{H} \left\langle \left(Q_{r,h}^{\pi^{k-1}} + Y^{k-1} Q_{g,h}^{\pi^{k-1}} \right)(x_h, \cdot), \pi_h(\cdot \mid x_h) \right\rangle$$

Lagrangian-based improvement

$$- \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h(\cdot \mid x_h), \tilde{\pi}_h^{k-1}(\cdot \mid x_h) \right)$$

Regularization

Achiam, Held, Tamar, & Abbeel, '17
Cai, Yang, Jin, Wang, '20

Primal-Dual Proximal Policy Optimization

- Primal update

$$\pi^k \leftarrow \underset{\pi}{\operatorname{argmax}} \sum_{h=1}^{H} \left\langle \left(Q_{r,h}^{\pi^{k-1}} + Y^{k-1} Q_{g,h}^{\pi^{k-1}} \right)(x_h, \cdot), \pi_h(\cdot \mid x_h) \right\rangle$$

Lagrangian-based improvement

$$- \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h(\cdot \mid x_h), \tilde{\pi}_h^{k-1}(\cdot \mid x_h) \right)$$

Regularization

- Dual update

$$Y^k \leftarrow \operatorname{Proj}\left(Y^{k-1} + \eta \left(b - V_{g,1}^{\pi^{k-1}}(x_1) \right) \right)$$

Achiam, Held, Tamar, & Abbeel, '17
Cai, Yang, Jin, Wang, '20

Primal-Dual Proximal Policy Optimization

➢ Primal policy update

$$\pi^k \leftarrow \underset{\pi}{\text{argmax}} \sum_{h=1}^{H} \left\langle \left(Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}\right)(x_h, \cdot), \pi_h(\cdot\,|x_h) \right\rangle$$

$$- \frac{1}{\alpha} \sum_{h=1}^{H} D\left(\pi_h(\cdot\,|x_h), \tilde{\pi}_h^{k-1}(\cdot\,|x_h)\right)$$

Lagrangian-based improvement

Regularization

➢ Dual update

$$Y^k \leftarrow \text{Proj}\left(Y^{k-1} + \eta\left(b - V_{g,1}^{k-1}(x_1)\right)\right)$$

Achiam, Held, Tamar, & Abbeel, '17
Cai, Yang, Jin, Wang, '20

Policy Evaluation With Optimism

- Upper confidence bound (UCB) exploration

$$Q^k_{r,h} \cong \underbrace{\varphi^T u^k_{r,h}}_{r_h} + \underbrace{\left(\phi^\tau_{r,h}\right)^T \omega^k_{r,h}}_{\mathbb{P}_h V^k_{r,h+1}} + \underbrace{\Gamma^k_h + \Gamma^k_{r,h}}_{\text{UCBs}} \geq Q^{\pi^k}_{r,h}$$

- Least-squares temporal difference

$$u^k_{r,h} \leftarrow \underset{u}{\text{argmin}} \sum_{\tau=1}^{k-1} \left(r_h(x^\tau_h, a^\tau_h) - \varphi(x^\tau_h, a^\tau_h)^T u \right)^2 + \lambda \|u\|^2$$

$$\omega^k_{r,h} \leftarrow \underset{\omega}{\text{argmin}} \sum_{\tau=1}^{k-1} \left(V^\tau_{r,h+1}(x^\tau_{h+1}) - \phi^\tau_{r,h}(x^\tau_h, a^\tau_h)^T \omega \right)^2 + \lambda \|\omega\|^2$$

Cai, Yang, Jin, & Wang, '99

Our Result

- Algorithm: Optimistic Primal-Dual Proximal Policy Optimization

  Primal-dual proximal policy optimization + Optimistic policy evaluation

Our Result

- Algorithm: Optimistic Primal-Dual Proximal Policy Optimization

  Primal-dual proximal policy optimization + Optimistic policy evaluation

- Regret and constraint violation guarantees

  $\text{Regret}(K) = \tilde{O}(d\, H^{2.5} \sqrt{T})$ $\text{Violation}(K) = \tilde{O}(d\, H^{2.5} \sqrt{T})$

  T: Total number of steps; H: Horizon length; d: Dimension of features.

  ✓ $d^2 H^5/\epsilon^2$ - polynomial sample complexity

  ✓ No any strong requirements on sampling models

  ✓ No explicit dependence on state space size $|\mathcal{S}|$

Poster Session 2

April 13 at 18:30-20:30 PDT

Thank you!