Model-based Offline RL with Partial Coverage

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: CPPO

1. MLE: $\hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$

2. Constrained Pessimistic Policy Optimization

Recap: CPPO

$$\text{1. MLE: } \hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$$

2. Constrained Pessimistic Policy Optimization

$$\max_{\pi} \min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \mid s, a) - \hat{P}(\cdot \mid s, a) \right\|_1^2 \leq \delta$$

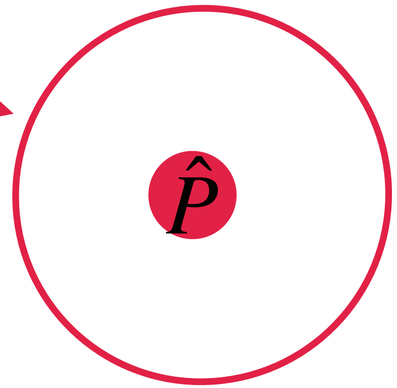
Recap: CPPO

1. MLE: $\hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$

2. Constrained Pessimistic Policy Optimization

$$\max_{\pi} \min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \mid s, a) - \hat{P}(\cdot \mid s, a) \right\|_1^2 \leq \delta$$



Recap: CPPO

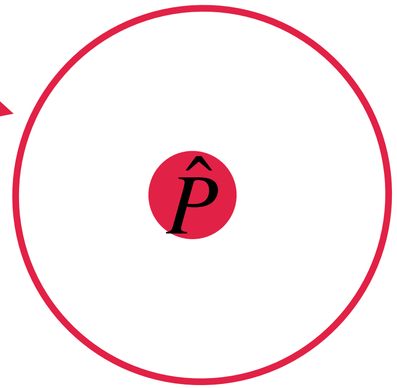
$$\text{1. MLE: } \hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$$

2. Constrained Pessimistic Policy Optimization

$$\max_{\pi} \min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \mid s, a) - \hat{P}(\cdot \mid s, a) \right\|_1^2 \leq \delta$$

$$\left(\text{or } \frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a) \geq \frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln \hat{P}(s' \mid s, a) - \delta \right)$$



Recap: CPPO

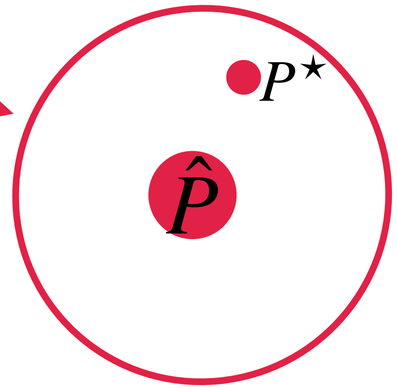
$$\text{1. MLE: } \hat{P} = \max_{P \in \mathcal{P}} \sum_{s,a,s' \in \mathcal{D}} \ln P(s' \mid s, a)$$

2. Constrained Pessimistic Policy Optimization

$$\max_{\pi} \min_{P \in \mathcal{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathcal{D}|} \sum_{s,a \in \mathcal{D}} \left\| P(\cdot \mid s, a) - \hat{P}(\cdot \mid s, a) \right\|_1^2 \leq \delta$$

$$\left(\text{or } \frac{1}{|\mathcal{D}|} \sum_{s,a,s' \in \mathcal{D}} \ln P(s' \mid s, a) \geq \frac{1}{|\mathcal{D}|} \sum_{s,a,s' \in \mathcal{D}} \ln \hat{P}(s' \mid s, a) - \delta \right)$$



Recap: CPPO

1. MLE: $\hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$

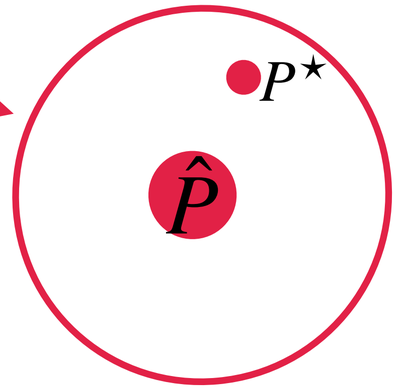
2. Constrained Pessimistic Policy Optimization

$$\max_{\pi} \min_{P \in \mathscr{P}} J(\pi; P)$$

Select the least favorable model!

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \mid s, a) - \hat{P}(\cdot \mid s, a) \right\|_1^2 \leq \delta$$

$$\left(\text{or } \frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a) \geq \frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln \hat{P}(s' \mid s, a) - \delta \right)$$



Formal Theoretical Guarantee for CPPO

2. CPPO's Sample Complexity:

Given n (i.i.d) offline data points, with high probability:

$$\forall \pi^*; V_{P\star}^{\pi^*} - V_{P\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{C_{\pi^*}^{\dagger} \ln(|\mathscr{P}|/\delta)}{n}}\right)$$

Formal Theoretical Guarantee for CPPO

2. CPPO's Sample Complexity:

Given n (i.i.d) offline data points, with high probability:

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{C_{\pi^*}^{\dagger} \ln(|\mathscr{P}|/\delta)}{n}}\right)$$

The cost we pay if want to compete w/ less covered policy π^*

Formal Theoretical Guarantee for CPPO

2. CPPO's Sample Complexity:

Given n (i.i.d) offline data points, with high probability:

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{C_{\pi^*}^{\dagger} \ln(|\mathcal{P}|/\delta)}{n}}\right)$$

The cost we pay if want to compete w/ less covered policy π^*

Robustness!

Formal Theoretical Guarantee for CPPO

2. CPPO's Sample Complexity:

Given n (i.i.d) offline data points, with high probability:

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{C_{\pi^*}^{\dagger} \ln(|\mathscr{P}|/\delta)}{n}}\right)$$

The cost we pay if want to compete w/ less covered policy π^*

Statistical complexity of $\mathscr{P}$; no poly dependence on $|S|, |A|$

Robustness!

Formal Theoretical Guarantee for CPPO

2. CPPO's Sample Complexity:

Given n (i.i.d) offline data points, with high probability:

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{C_{\pi^*}^{\dagger} \ln(|\mathscr{P}|/\delta)}{n}}\right)$$

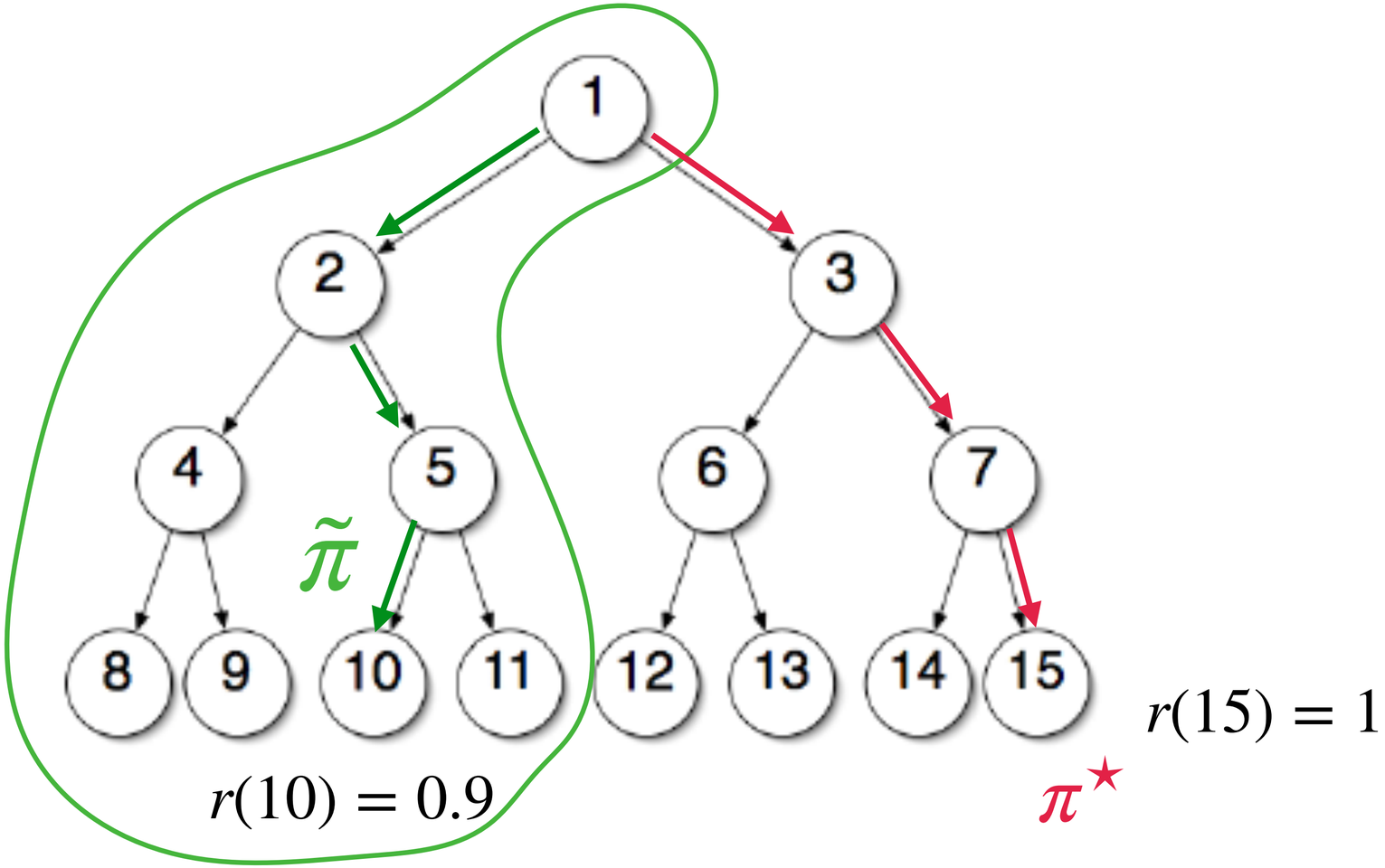
The cost we pay if want to compete w/ less covered policy π^*

Robustness!

Statistical complexity of $\mathscr{P}$; no poly dependence on $|S|, |A|$

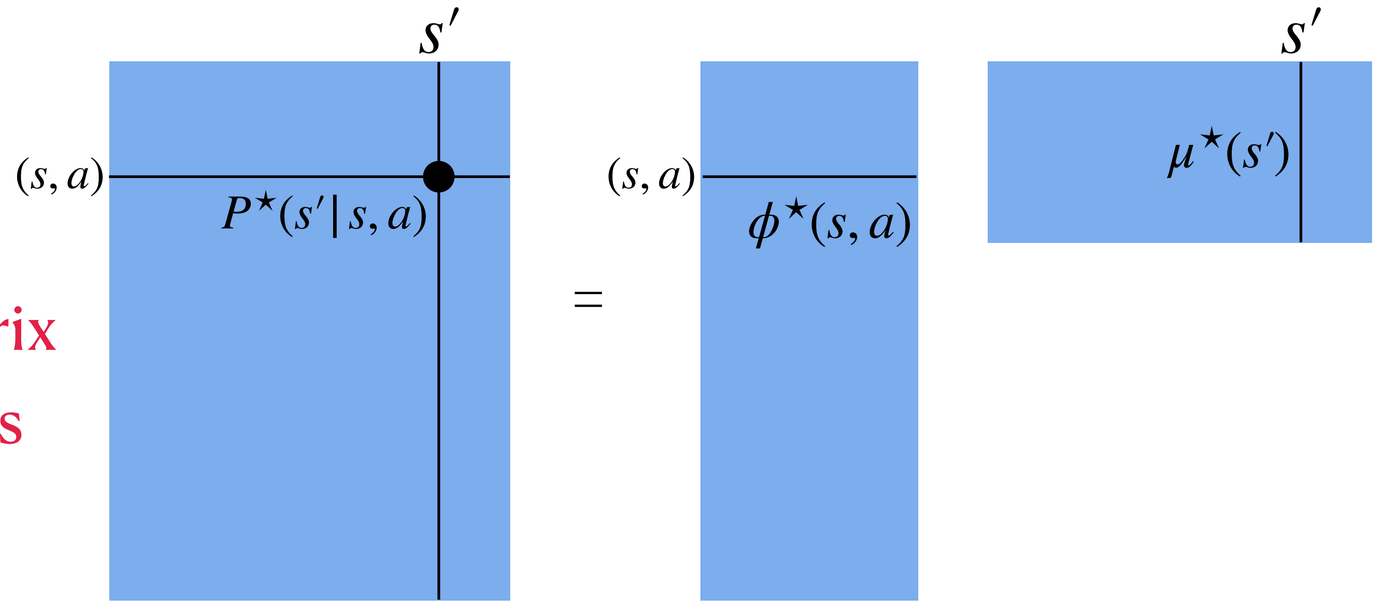
SL-style Generalization!

Would CPPO succeed here?



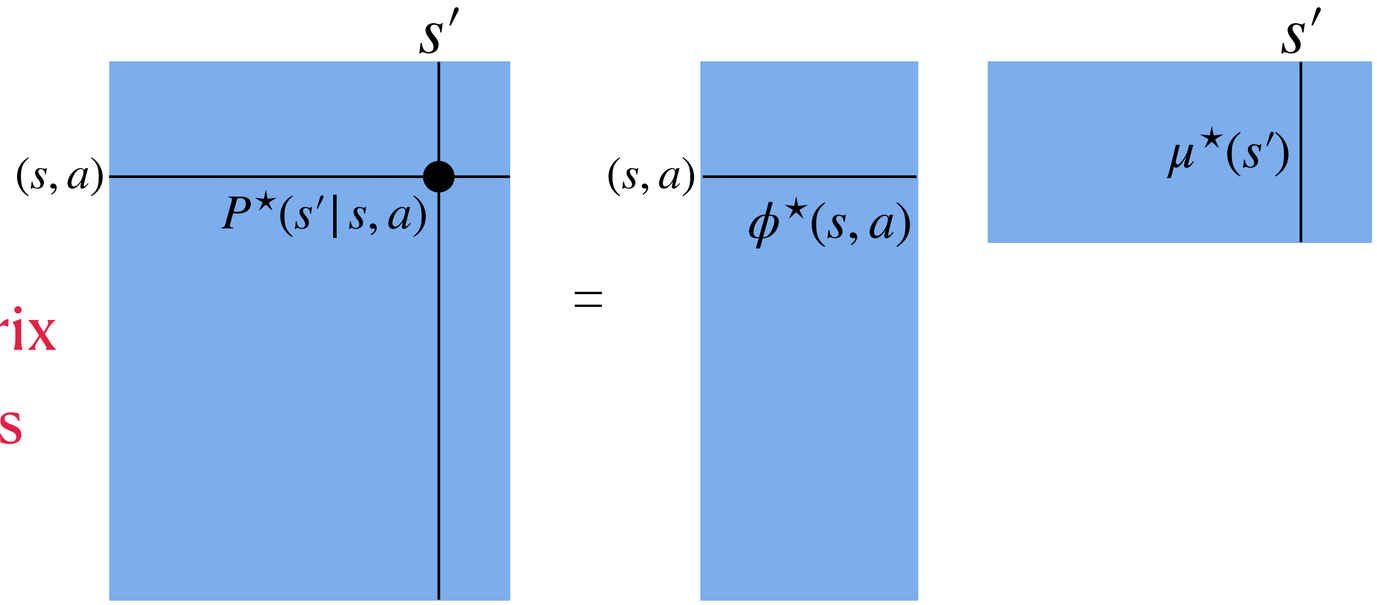
CPPO for Low-rank MDPs

Transition matrix $P \in \mathbb{R}^{SA \times S}$ has rank d



CPPO for Low-rank MDPs

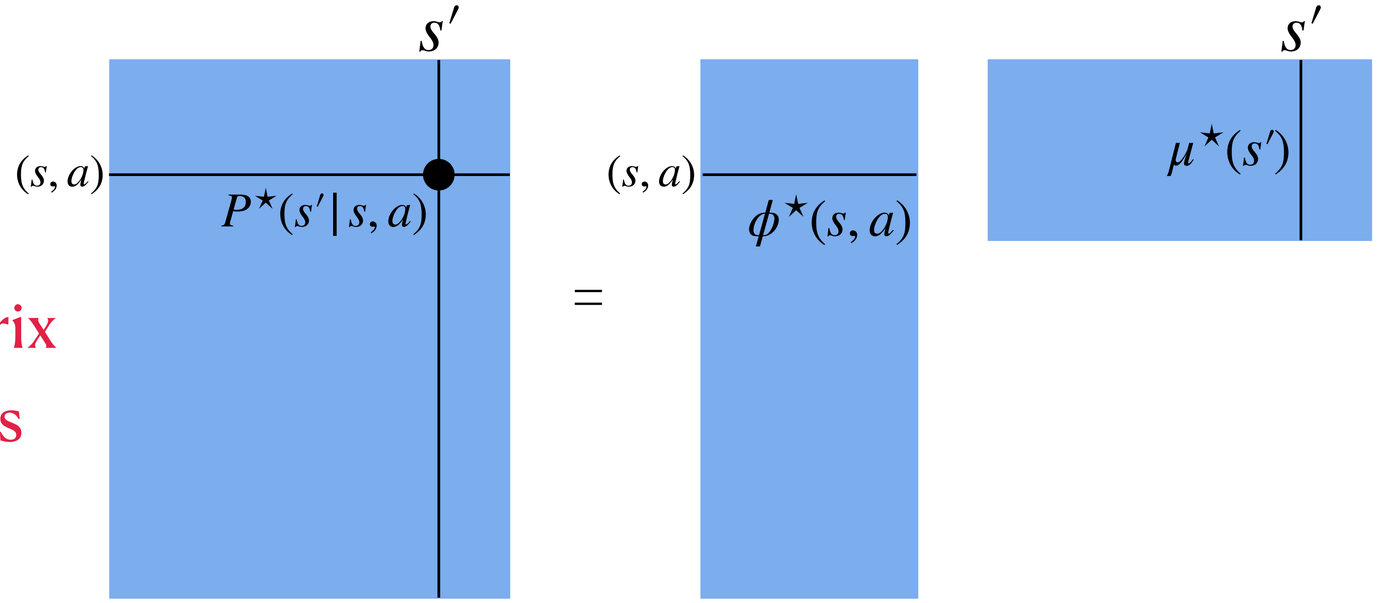
Transition matrix $P \in \mathbb{R}^{SA \times S}$ has rank d



$$\exists \mu^\star, \phi^\star : \quad \forall s, a, s', P^\star(s' \mid s, a) = \mu^\star(s')^\top \phi^\star(s, a)$$

CPPO for Low-rank MDPs

Transition matrix $P \in \mathbb{R}^{SA \times S}$ has rank d



$$\exists \mu^\star, \phi^\star : \quad \forall s, a, s', P^\star(s' \mid s, a) = \mu^\star(s')^\top \phi^\star(s, a)$$

In low-rank MDP, neither $\mu^\star$ nor $\phi^\star$ is known

CPPO for Low-rank MDPs

Two Function classes Γ & Φ

Realizability: $\mu^\star \in \Gamma$, $\phi^\star \in \Phi$

CPPO for Low-rank MDPs

Two Function classes Γ & Φ

Realizability: $\mu^\star \in \Gamma$, $\phi^\star \in \Phi$

Coverage def: relative condition number under ground truth (unknown) $\phi^\star$

$$C_\pi^\dagger = \max_x \frac{x^\top \mathbb{E}_{s,a\sim d^\pi} \phi^\star(s,a)\phi^\star(s,a)^\top x}{x^\top \mathbb{E}_{s,a\sim d^{\pi_b}} \phi^\star(s,a)\phi^\star(s,a)^\top x}$$

CPPO for Low-rank MDPs

Two Function classes Γ & Φ

Realizability: $\mu^\star \in \Gamma$, $\phi^\star \in \Phi$

Coverage def: relative condition number under ground truth (unknown) $\phi^\star$

$$C_\pi^\dagger = \max_x \frac{x^\top \mathbb{E}_{s,a\sim d^\pi} \phi^\star(s,a)\phi^\star(s,a)^\top x}{x^\top \mathbb{E}_{s,a\sim d^{\pi_b}} \phi^\star(s,a)\phi^\star(s,a)^\top x}$$

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{dC_{\pi^*}^\dagger \ln(|\Gamma||\Phi|/\delta)}{n}}\right)$$

CPPO for Low-rank MDPs

Two Function classes Γ & Φ

Realizability: $\mu^\star \in \Gamma$, $\phi^\star \in \Phi$

Coverage def: relative condition number under ground truth (unknown) $\phi^\star$

$$C_\pi^\dagger = \max_x \frac{x^\top \mathbb{E}_{s,a\sim d^\pi} \phi^\star(s,a)\phi^\star(s,a)^\top x}{x^\top \mathbb{E}_{s,a\sim d^{\pi_b}} \phi^\star(s,a)\phi^\star(s,a)^\top x}$$

$$\forall \pi^*; V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = O\left(H^2 \sqrt{\frac{dC_{\pi^*}^\dagger \ln(|\Gamma||\Phi|/\delta)}{n}}\right)$$

(Many more interesting examples: linear mixture MDP, factored MDP, etc)

Implementation

1. MLE: $\hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$

2: Treat constraint as a penalty w/ Lagrangian multiplier:

Implementation

$$1.\ \text{MLE: } \hat{P} = \max_{P \in \mathscr{P}} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a)$$

2: Treat constraint as a penalty w/ Lagrangian multiplier:

$$\max_{\pi} \min_{P} J(\pi; P) + \max_{\lambda \leq 0} \lambda \left(\frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln P(s' \mid s, a) - \frac{1}{|\mathscr{D}|} \sum_{s,a,s' \in \mathscr{D}} \ln \hat{P}(s' \mid s, a) + \delta \right)$$

Practical version of CPPO (Rigter et al. Neurips22)

"Uehara et al. (2021) provides the theoretical motivation for solving Problem 1. In this work, we focus on developing a practical approach to solving Problem 1."

Practical version of CPPO

CPPO

		Ours	Model-based baselines				Model-free baselines		
		RAMBO	**RepB-SDE**	**COMBO**	**MOPO**	**MOReL**	**CQL**	**IQL**	**TD3+BC**
Random	HalfCheetah	40.0 ± 2.3	32.9	38.8	35.4	25.6	19.6	-	11.0
	Hopper	21.6 ± 8.0	8.6	17.9	4.1	53.6	6.7	-	8.5
	Walker2D	11.5 ± 10.5	21.1	7.0	4.2	37.3	2.4	-	1.6
Medium	HalfCheetah	77.6 ± 1.5	49.1	54.2	69.5	42.1	49.0	47.4	48.3
	Hopper	92.8 ± 6.0	34.0	94.9	48.0	95.4	66.6	66.3	59.3
	Walker2D	86.9 ± 2.7	72.1	75.5	-0.2	77.8	83.8	78.3	83.7
Medium Replay	HalfCheetah	68.9 ± 2.3	57.5	55.1	68.2	40.2	47.1	44.2	44.6
	Hopper	96.6 ± 7.0	62.2	73.1	39.1	93.6	97.0	94.7	60.9
	Walker2D	85.0 ± 15.0	49.8	56.0	69.4	49.8	88.2	73.9	81.8
Medium Expert	HalfCheetah	93.7 ± 10.5	55.4	90.0	72.7	53.3	90.8	86.7	90.7
	Hopper	83.3 ± 9.1	82.6	111.1	3.3	108.7	106.8	91.5	98.0
	Walker2D	68.3 ± 20.6	88.8	96.1	-0.3	95.6	109.4	109.6	110.1
MuJoCo-v2 Total:		826.2 ± 33.8	614.1	769.7	413.4	773.0	767.4	692.6*	698.5

SOTA

Proof sketch for CPPO

1. MLE generalization bound

Given a dataset $\mathscr{D} = \{x^i, y^i\}_{i=1}^N$, where $y^i \sim P^\star(\cdot \mid x_i)$, denote $\hat{P}$ as the MLE, i.e.,

$$\hat{P} = \arg\max_{P \in \mathscr{P}} \sum_{i=1}^{N} \ln P(y_i \mid x_i)$$

Proof sketch for CPPO

1. MLE generalization bound

Given a dataset $\mathscr{D} = \{x^i, y^i\}_{i=1}^N$, where $y^i \sim P^\star(\cdot \mid x_i)$, denote $\hat{P}$ as the MLE, i.e.,

$$\hat{P} = \arg\max_{P \in \mathscr{P}} \sum_{i=1}^{N} \ln P(y_i \mid x_i)$$

assuming $P^\star \in \mathscr{P}$, then with prob $1-\delta$, we have: $\frac{1}{N} \sum_{i=1}^{N} \|\hat{P}(\cdot \mid x_i) - P^\star(\cdot \mid x_i)\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$

Proof sketch for CPPO

1. MLE generalization bound

Given a dataset $\mathscr{D} = \{x^i, y^i\}_{i=1}^N$, where $y^i \sim P^\star(\,\cdot\,|\,x_i)$, denote $\hat{P}$ as the MLE, i.e.,

$$\hat{P} = \arg\max_{P \in \mathscr{P}} \sum_{i=1}^{N} \ln P(y_i \,|\, x_i)$$

assuming $P^\star \in \mathscr{P}$, then with prob $1-\delta$, we have: $\frac{1}{N}\sum_{i=1}^{N} \|\hat{P}(\,\cdot\,|\,x_i) - P^\star(\,\cdot\,|\,x_i)\|_1^2 \le \frac{\ln(|\,\mathscr{P}\,|/\delta)}{N}$

If $x_i \sim_{i.i.d} \nu$, then: $\mathbb{E}_{x\sim\nu}\|\hat{P}(\,\cdot\,|\,x) - P^\star(\,\cdot\,|\,x)\|_1^2 \le \frac{\ln(|\,\mathscr{P}\,|/\delta)}{N}$

Proof sketch for CPPO

2. $P^\star$ is always a feasible solution of our constrainted optimization problems

$$\max_{\pi} \min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Claim 1: For all π, denote $\underline{V}^{\pi} = \min_{P \in \mathscr{P}_{\mathscr{D}}} V_P^{\pi}$, we have $\underline{V}^{\pi} \leq V_{P^\star}^{\pi}$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Claim 1: For all π, denote $\underline{V}^{\pi} = \min_{P \in \mathscr{P}_{\mathscr{D}}} V_P^{\pi}$, we have $\underline{V}^{\pi} \leq V_{P^\star}^{\pi}$

Claim 2: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Claim 1: For all π, denote $\underline{V}^{\pi} = \min_{P \in \mathscr{P}_{\mathscr{D}}} V_P^{\pi}$, we have $\underline{V}^{\pi} \leq V_{P^\star}^{\pi}$

Claim 2: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$

Proof: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*} + \underline{V}^{\pi^*} - V_{P^\star}^{\hat{\pi}}$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Claim 1: For all π, denote $\underline{V}^{\pi} = \min_{P \in \mathscr{P}_{\mathscr{D}}} V_P^{\pi}$, we have $\underline{V}^{\pi} \leq V_{P^\star}^{\pi}$

Claim 2: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$

Proof: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*} + \underline{V}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*} + \underline{V}^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}}$

Proof sketch for CPPO

3. Pessimism

$$\left(\hat{\pi}, \hat{P}\right) = \arg\max_{\pi} \arg\min_{P \in \mathscr{P}} J(\pi; P)$$

$$\text{s.t.,} \frac{1}{|\mathscr{D}|} \sum_{s,a \in \mathscr{D}} \left\| P(\cdot \,|\, s,a) - \hat{P}(\cdot \,|\, s,a) \right\|_1^2 \leq \frac{\ln(|\mathscr{P}|/\delta)}{N}$$

Claim 1: For all π, denote $\underline{V}^{\pi} = \min_{P \in \mathscr{P}_{\mathscr{D}}} V_P^{\pi}$, we have $\underline{V}^{\pi} \leq V_{P^\star}^{\pi}$

Claim 2: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$

Proof: $V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} = V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*} + \underline{V}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*} + \underline{V}^{\hat{\pi}} - V_{P^\star}^{\hat{\pi}}$

$$\leq V_{P^\star}^{\pi^*} - \underline{V}^{\pi^*}$$

Proof sketch for CPPO

3. Final step: simulation lemma and distribution change

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$$

Proof sketch for CPPO

3. Final step: simulation lemma and distribution change

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$$

$$\leq H^2 \mathbb{E}_{s,a\sim d^{\pi^*}} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$$

Proof sketch for CPPO

3. Final step: simulation lemma and distribution change

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$$

$$\leq H^2 \mathbb{E}_{s,a\sim d^{\pi^*}} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$$

$$\leq H^2 \sqrt{\mathbb{E}_{s,a\sim d^{\pi^*}} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2}$$

Proof sketch for CPPO

3. Final step: simulation lemma and distribution change

$$V_{P^\star}^{\pi^*} - V_{P^\star}^{\hat{\pi}} \leq V_{P^\star}^{\pi^*} - V_{\hat{P}}^{\pi^*}$$

$$\leq H^2 \mathbb{E}_{s,a\sim d^{\pi^*}} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1$$

$$\leq H^2 \sqrt{\mathbb{E}_{s,a\sim d^{\pi^*}} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2}$$

$$\leq H^2 \sqrt{C^{\pi^*} \mathbb{E}_{s,a\sim \nu} \|\hat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_1^2}$$