A Provably Better Offline RL Algorithm

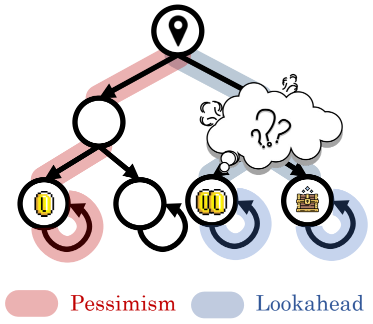
Don't be Pessimistic Too Early: Look K Steps Ahead!

Chaoqi Wang
University of Chicago

Ziyu Ye
University of Chicago

Kevin Murphy
Google Research

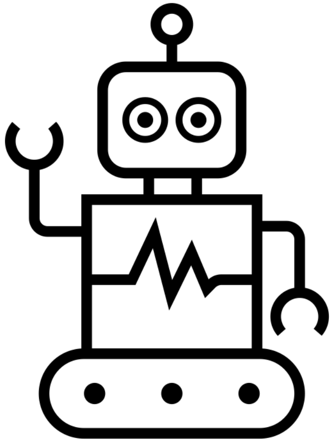
Yuxin Chen
University of Chicago



February 2023

Agenda

1. Introduction
    - Offline RL & the Pessimism Principle
    - The Excessive Pessimism Dilemma
    - Our Contributions
2. Methods
    - Formal statement
    - Lookahead Pessimistic MDP (LP-MDP)
3. Theoretical Analysis
    - The Suboptimality Bound
    - The Effect of Lookahead Horizon
4. Experimental Results
5. Takeaways and Future Works



Part 1
Introduction

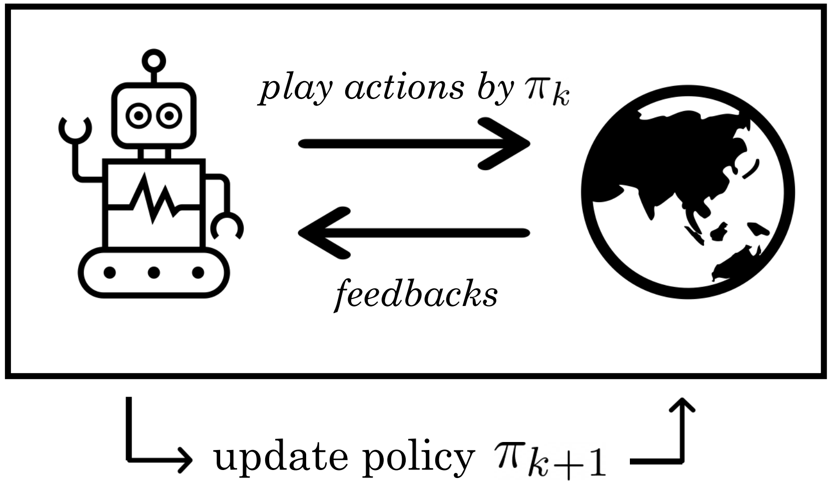
Reinforcement Learning (RL)

RL is about training agents to **learn** to **make sequential decisions** to achieve **goals**.

Reinforcement Learning (RL)

RL is about training agents to **learn** to **make sequential decisions** to achieve **goals**.

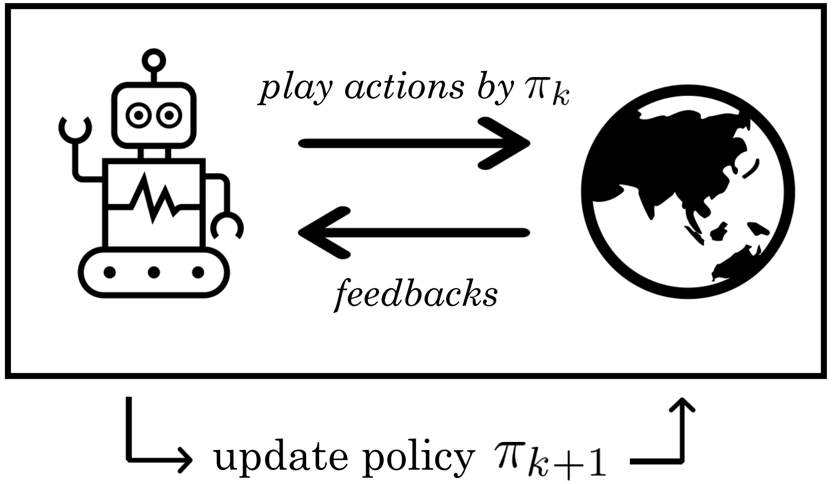
Classical RL is Online



Reinforcement Learning (RL)

RL is about training agents to **learn** to **make sequential decisions** to achieve **goals**.

Classical RL is Online



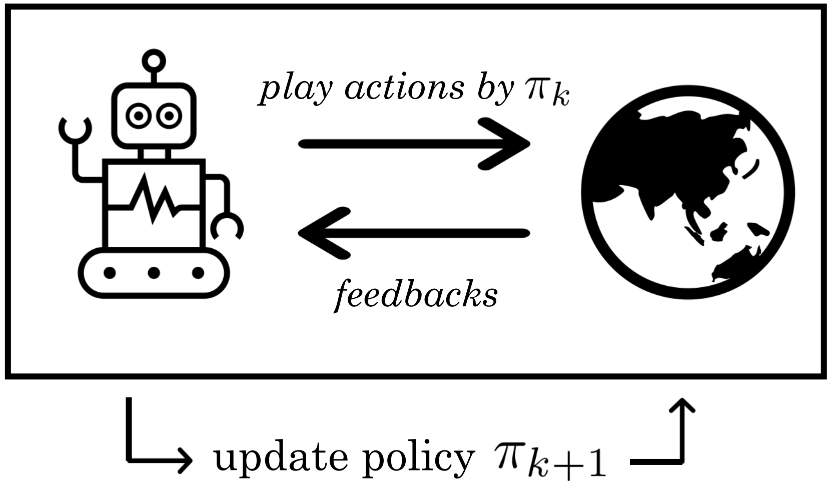
requires online interactions

unsafe time-consuming costly

Reinforcement Learning (RL)

RL is about training agents to **learn** to **make sequential decisions** to achieve **goals**.

Classical RL is Online



⚠ requires online explorations

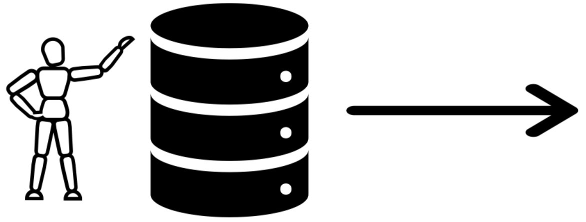
unsafe time-consuming costly

In reality, we often have massive pre-collected data (e.g., by human demonstration).

Can we still train RL policies *without* any online explorations?

Offline RL

Offline RL



learn policy π

a static dataset of trajectories
(pre-collected by any means)

- No environment interactions
- No further data collections

Formally...

$$\mathcal{D} = \{(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_i)\}$$
$$\mathbf{s} \sim d^{\pi_\beta}(\mathbf{s})$$
$$\mathbf{a} \sim \pi_\beta(\mathbf{a} \mid \mathbf{s})$$
$$\mathbf{s}' \sim p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$$
$$r \leftarrow r(\mathbf{s}, \mathbf{a})$$

Offline RL Objective

$$\max_\pi \sum_{t=0}^{T} E_{\mathbf{s}_t \sim d^\pi(\mathbf{s}), \mathbf{a}_t \sim \pi(\mathbf{a}|\mathbf{s})} \left[\gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$$

Offline RL: *Challenges*

Offline RL

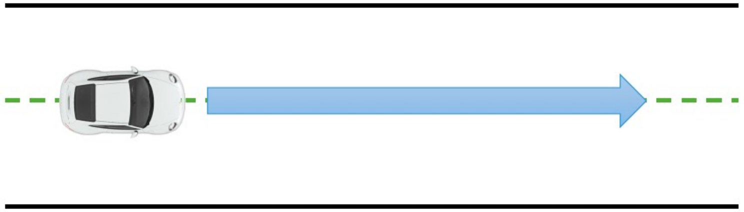
learn policy π

a static dataset of trajectories
(pre-collected by any means)

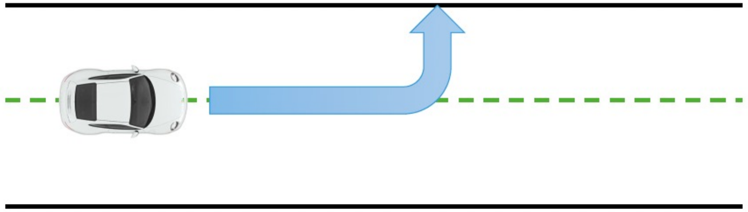
- No environment interactions
- No further data interaction

Fundamental challenge: *erroneous extrapolation.*

Training data only contains:



What the policy wants to do…



Good?
Bad?

Online RL can tackle this by trial & error.

How may offline RL deal with potential **out-of-distribution** actions?

The Pessimism Principle in the Face of Uncertainty

Key idea: *avoiding uncertain state & actions* by explicit penalization.

- Wang et al. (2020) regularize the learned policy.
- Kostrikov et al. (2022) penalize the rewards..
- Kidambi et al. (2020) truncate transitions.

The Pessimism Principle in the Face of Uncertainty

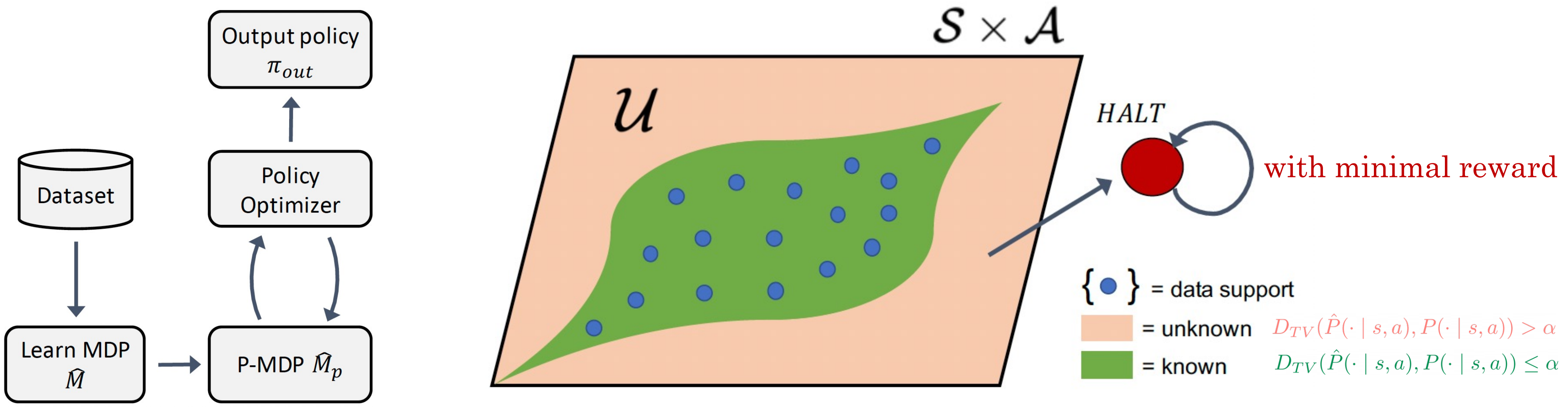


Figure 1. An illustration of Pessimistic Markov Decision Process (P-MDP) by Kidambi et al. (2020).
Notice that the **value** of any policy in P-MDP will be the **lower bound** for the true value

Key idea: *avoiding uncertain state & actions* by explicit penalization.

- Wang et al. (2020) regularize the learned policy.
- Kostrikov et al. (2022) penalize the rewards..
- Kidambi et al. (2020) truncate transitions.

What can go wrong for pessimism based algorithms?

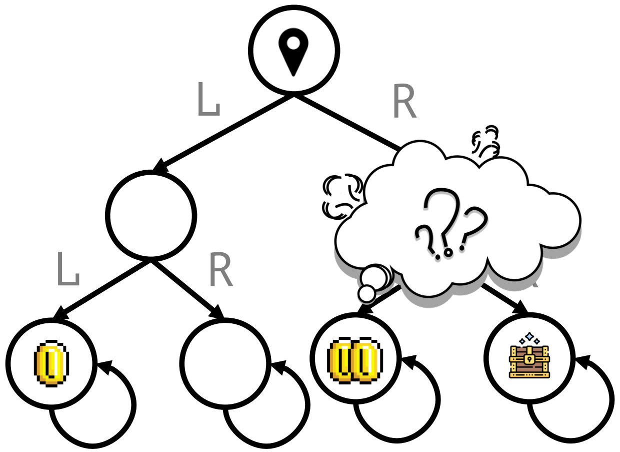
The *Excessive* Pessimism Dilemma

The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.

The *Excessive Pessimism* Dilemma

The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.

(Fitted MDP)

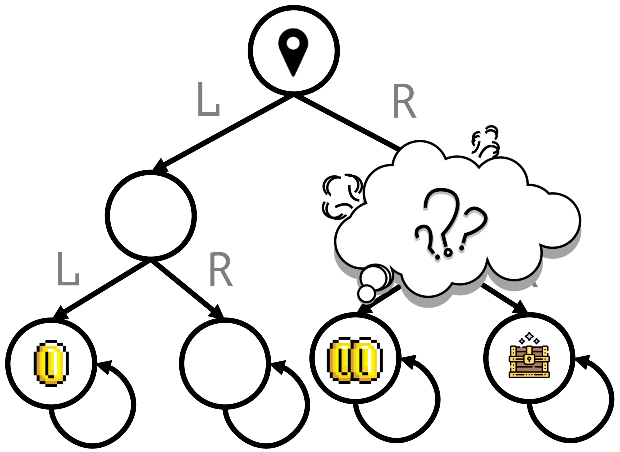


The *Excessive* Pessimism Dilemma

⚠ The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.

$r(\ ) = 1$ $r(\ ) = 2$ $r(\ ) = 10$ Uncertain region

(Fitted MDP)

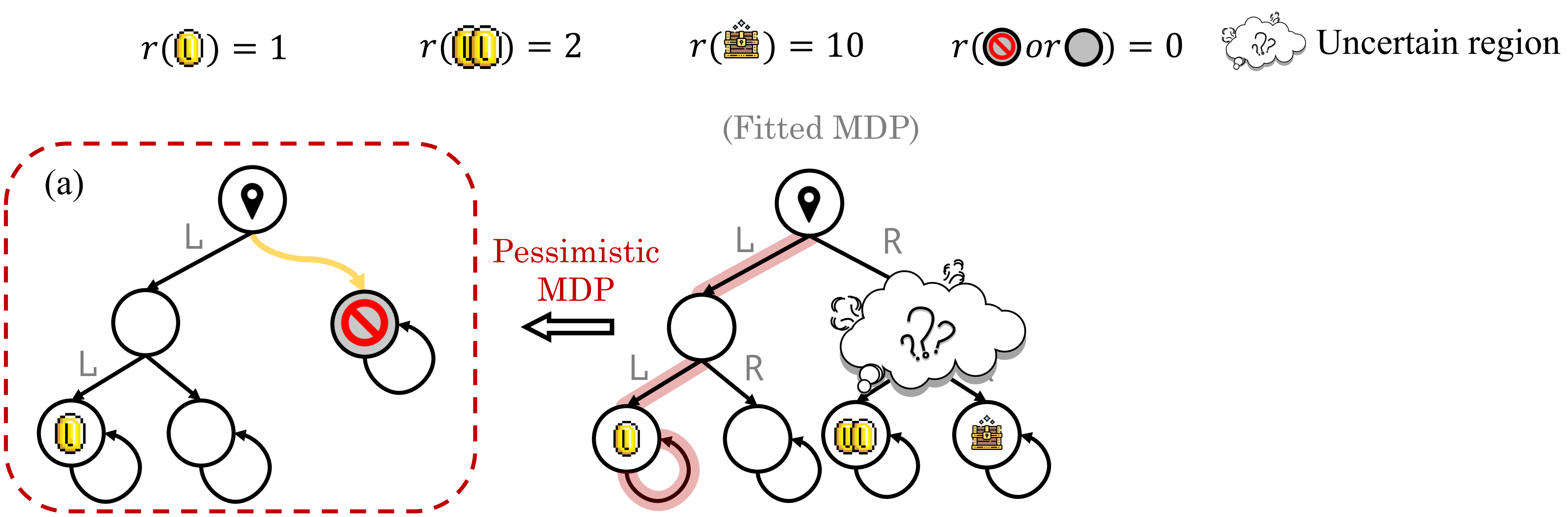


Classical Pessimistic MDP will directly halts here.

No chance to get > !

The *Excessive* Pessimism Dilemma

⚠ The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.



$\pi^*_{\text{P-MDP}} = (\text{L}, \text{L})$

Halting too early → get at best .

Possible Path of $\pi^*_{\text{P-MDP}}$

Can we find a principled way to *modulate* pessimism?

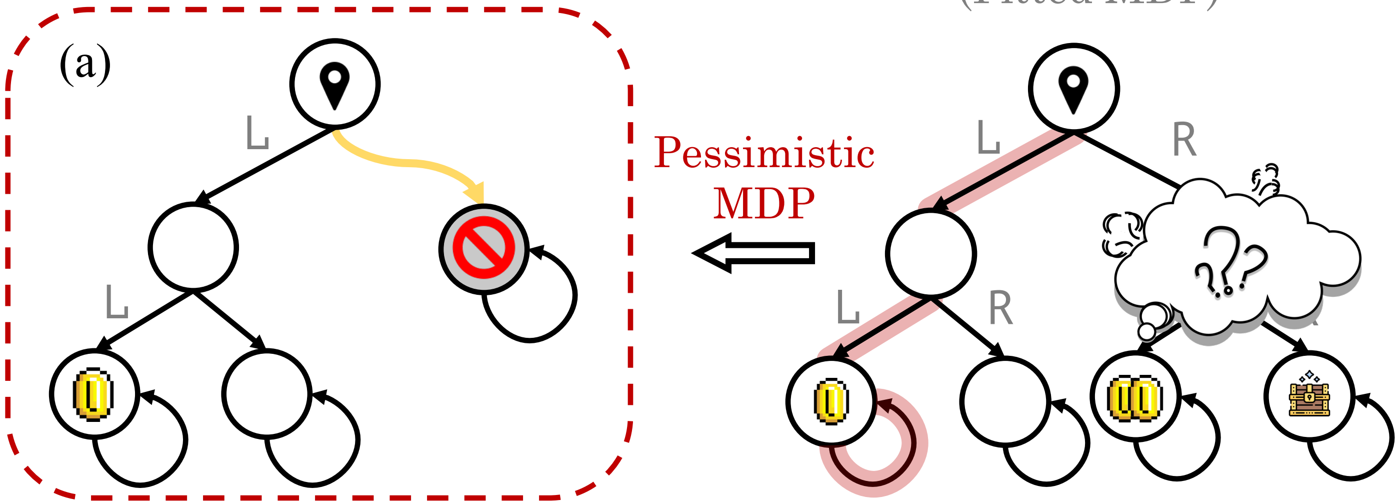
(And to achieve a better performance guarantee…)

The *Excessive* Pessimism Dilemma: *Mitigating* by *Lookahead*

The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.

$r(\ ) = 1$ $r(\ ) = 2$ $r(\ ) = 10$ $r(\ or\ ) = 0$ Uncertain region

(Fitted MDP)



Just look one step ahead…

Fortunes (or) await for sure!

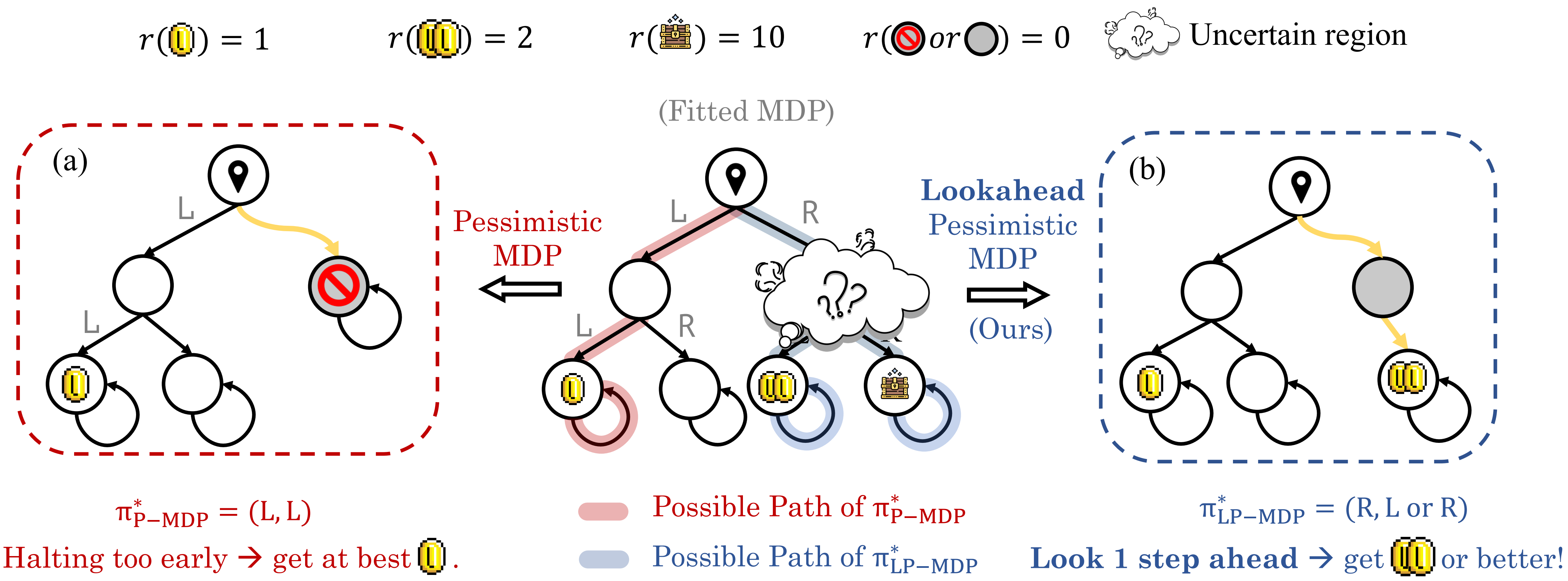
$\pi^*_{\text{P-MDP}} = (\text{L}, \text{L})$

Halting too early → get at best .

Possible Path of $\pi^*_{\text{P-MDP}}$

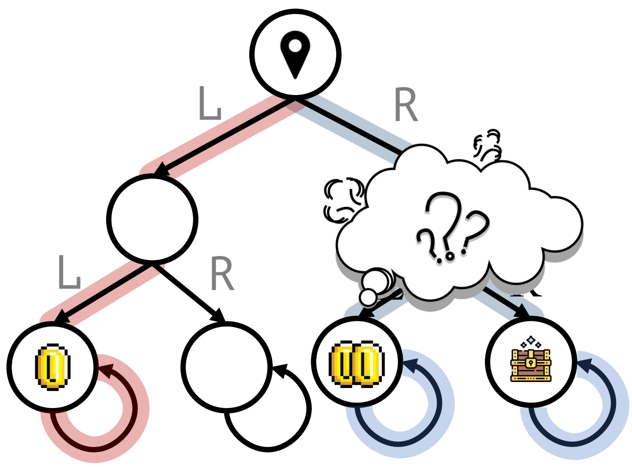
The *Excessive Pessimism* Dilemma: *Mitigating by Lookahead*

⚠ The policy may *behave overly conservative*, ends up too *far away* from achievable better performance.



Lookahead enables a *less conservative* policy with *better performance guarantee*!

Further Insights on Lookahead Pessimism



Possible Path of π^*_{P-MDP}

Possible Path of π^*_{LP-MDP}

- The lookahead horizon **modulates the pessimism level**.
- Lookahead helps circumvent uncertain areas by **path stitching**.
- Lookahead **implicitly increases the data coverage**.

Our Contributions

Algorithm: Lookahead Pessimistic MDP.

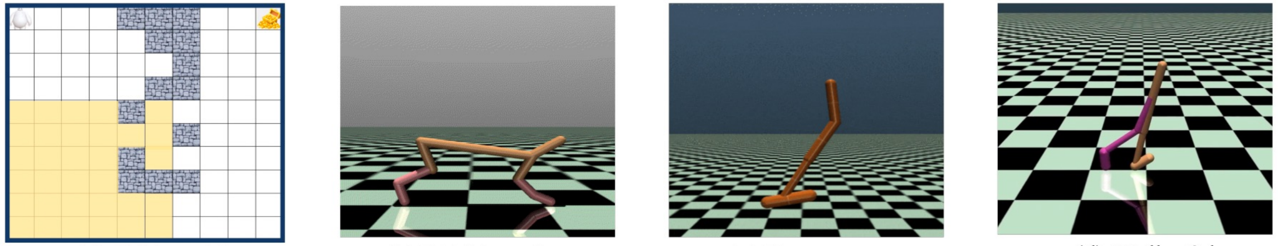
Theory: a lower bound monotonically improves with the lookahead horizon K.

Experiments: solid improvement over baselines on benchmark environments.

Algorithm 1 Policy Learning under Pessimistic MDP with look-ahead (LP(ξ, π, K)-MDP).

Require: Offline data $\mathcal{D}$, threshold ξ, look-ahead steps K, and learning rate α.

Fit the dynamics model $\mathcal{M}_{\widehat{p}}$ on $\mathcal{D}$
Initialize the policy π_0
for $n = 0, 1, 2, \ldots$ **do**
  Construct the LP(ξ, π, K)-MDP for π_n: $\mathcal{M}_{\widetilde{p}}^{\pi_n}$
  Improve policy: $\pi_{n+1} \leftarrow \texttt{SAC\_Step}(\pi_n, \mathcal{M}_{\widetilde{p}}^{\pi_n})$
end for
return π_n



(a) Grid world (b) Halfcheetah (c) Hopper (d) Walker2d

Dataset	Environment	LP-MDP (Ours)	MOReL (SAC)	MOReL (NPG)	MOPO	CQL	SAC-Off	BEAR	BRAC-p	BRAC-v
medium	halfcheetah	42.6±5.5	43.4	42.1	54.2	44.4	-4.3	41.7	43.8	46.3
medium	hopper	101.9 ±1.1*	75.8	95.4	28.0	86.6	0.8	52.1	32.7	31.1
medium	walker2d	64.5±5.1	76.8	77.8	17.8	74.5	0.9	59.1	77.5	81.1
medium-replay	halfcheetah	48.5±2.1*	43.4	40.2	53.1	46.2	-2.4	38.6	45.4	47.7
medium-replay	hopper	101.2 ±0.8	101.1	93.6	67.5	48.6	3.5	33.7	0.6	0.6
medium-replay	walker2d	82.7 ±5.9*	46.5	49.8	39.0	32.6	1.9	19.2	-0.3	0.9
medium-expert	halfcheetah	51.1±0.9*	41.6	53.3	63.3	62.4	1.8	53.4	44.2	41.9
medium-expert	hopper	103.7±1.2*	78.5	108.7	23.7	111.0	1.6	96.3	1.9	0.8
medium-expert	walker2d	81.5±8.5*	68.0	95.6	44.6	98.7	-0.1	40.1	76.9	81.6
Average Scores		677.7 *	575.1	656.5	391.2	605.0	3.7	434.2	328.1	332.0

Table 1. Results on D4RL. We report the averaged normalized scores of 3 different random seeds with one standard errors. We highlight the best averaged scores by a blue box . Also, we use * to indicate the tasks where our method has a solid improvement over MOReL-SAC.

Pessimistic MDP with K-step Lookahead

Definition 3 (LP(ξ, π, K)-MDP) *For any (s_t, a_t), and $\mathcal{M}_{\widehat{p}}$ and $\mathcal{M}_p$, the LP(ξ, π, K)-MDP of $\mathcal{M}_p$, (i.e., $\mathcal{M}_{\widetilde{p}}^{\pi}$) is constructed by modifying the transition $\widehat{p}$ to be $\widetilde{p}$ as:*[a]

Case 1 *(current transition is certain) If $(s_t, a_t) \notin \mathcal{U}$, then*

$$\widetilde{p}(\cdot|s_t, a_t) = \widehat{p}(\cdot|s_t, a_t), \tag{9}$$

Case 2 *(current transition is uncertain) If $(s_t, a_t) \in \mathcal{U}$, then*

Case 2.1 *(all K-step look ahead is uncertain) If $(s_t, a_t) \in \mathcal{U}^{\pi}_{-(1:K)}$, then*

$$\widetilde{p}(s_{t+1} = \mathbf{e}|s_t, a_t) = 1, \tag{10}$$

Case 2.2 *(some k_{th}-step look ahead is certain) If $(s_t, a_t) \in \mathcal{U}^{\pi}_{k}$ for some $k \in [K]$, then we construct a deterministic path such that $\forall\, i \in \{1, \ldots, k-1\}$,*

$$\widetilde{p}(s_{t+k} = s^{\star}|s_t, a_t, \pi) = 1,\ \widetilde{r}(s_{t+i}, \cdot) = -R_{\max}. \tag{11}$$

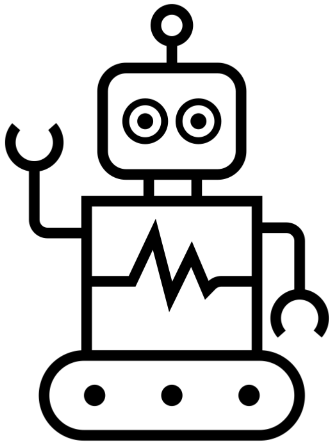
where $s^{\star}$ is defined as:

$$s^{\star} = \mathop{\arg\min}_{s' \in \mathcal{L}^{\pi,k}_{\mathcal{M}_{\widehat{p}}}(s_t, a_t)} V^{\pi}_{\mathcal{M}^{\pi}_{\widetilde{p}}}(s'), \tag{12}$$

[a] The associated reward $\widetilde{r}(\cdot) := r(\cdot)$ unless otherwise stated.

We are all in the gutter, but some of us are *looking at the stars*.

— *Oscar Wilde*



Part 2

Methodology

Preliminaries

Intuition

If the current state-action pair is *uncertain* → don't just halt!

Look a few steps ahead → if future states is promising w/ *high certainty*, go for it!

Preliminaries

Intuition

If the current state-action pair is *uncertain* → don't just halt!

Look a few steps ahead → if future states is promising w/ *high certainty*, go for it!

Uncertainty Quantification

We first denote the *estimation error* on any pair $(\boldsymbol{s}, \boldsymbol{a})$ as

$$d(\boldsymbol{s}, \boldsymbol{a}) := \mathbf{d}_{\mathrm{TV}}\left(\widehat{p}(\cdot|\boldsymbol{s}, \boldsymbol{a}) \| p(\cdot|\boldsymbol{s}, \boldsymbol{a})\right),$$

which quantifies the total variation distance from the estimated distribution $\widehat{p}$ to the true distribution p.

Preliminaries

Intuition

If the current state-action pair is *uncertain* → don't just halt!

Look a few steps ahead → if future states is promising w/ *high certainty*, go for it!

Uncertainty Quantification

We first denote the *estimation error* on any pair $(\boldsymbol{s}, \boldsymbol{a})$ as

$$d(\boldsymbol{s}, \boldsymbol{a}) := \mathbf{d}_{\mathrm{TV}}\left(\widehat{p}(\cdot|\boldsymbol{s}, \boldsymbol{a}) | \, p(\cdot|\boldsymbol{s}, \boldsymbol{a})\right),$$

which quantifies the total variation distance from the estimated distribution $\widehat{p}$ to the true distribution p.

Classical Pessimism asks to halt at …

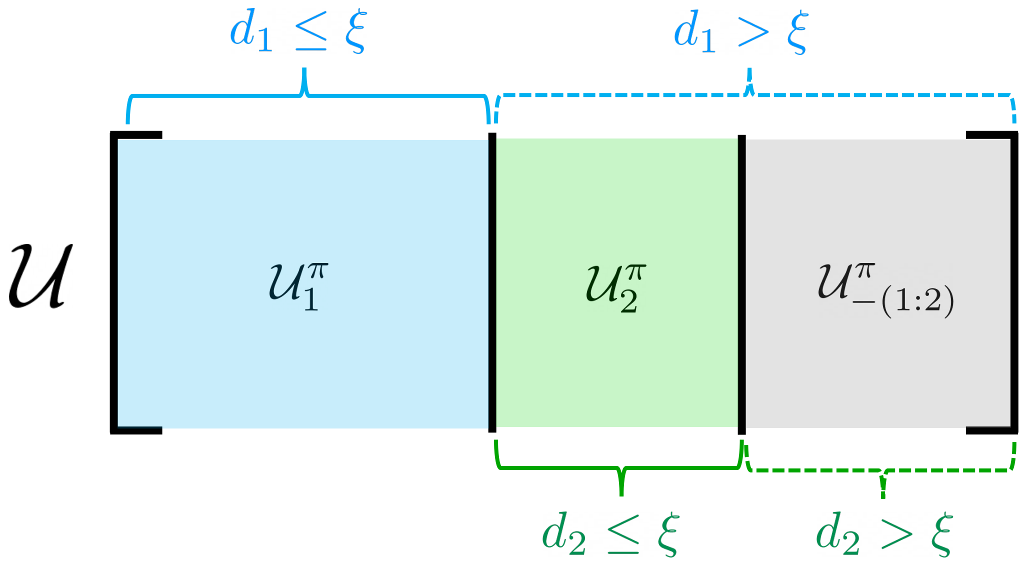
Definition 1 (Uncertain State-Action Set) $\forall\ \xi \geq 0$, *the set of uncertain state-action pairs is*

$$\mathcal{U}(\xi) := \{(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A} : d(\boldsymbol{s}, \boldsymbol{a}) \geq \xi\}\,. \qquad (7)$$

For simplicity, we drop the dependency on ξ in the notations.

Partition Uncertain Regions

We can further **partition** U by properties of some **lookahead sets**.



Associate Each Pair with Lookahead Sets

Definition 2 (Lookahead Set) *For any $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$ under MDP $\mathcal{M}_{p'}$ with $p'(\cdot|\cdot,\cdot,\pi)$ denoting the transition distribution relying on π, the k_{th}-step lookahead set is*

$$\mathcal{L}^{\pi,k}_{\mathcal{M}_{p'}}(s_t, a_t) := \{s \in \mathcal{S} |\ p'(s_{t+k} = s | s_t, a_t, \pi) > 0\}.$$

Partition Criteria: Lookahead Certainty

We say a state-action pair (s, a) is **k_{th}-certain** if for all the states in its k_{th}-step lookahead sets, the fitted dynamics $\mathcal{M}_{\widehat{p}}$ induces only a small estimation error, that is:

$$\forall s' \in \mathcal{L}^{\pi,k}_{\mathcal{M}_{\widehat{p}}}(s, a),\ d(s', \pi(s')) \leq \xi, \tag{8}$$

where $d(s, \pi(s)) := \max_{a:\pi(a|s)>0} d(s, a)$.

The set $\mathcal{U}$ thus can be partitioned into disjoint subsets by the above lookahead certainty criteria. For all $k \in [1, K]$, we define the subset $\mathcal{U}_k^\pi$ as

$$\mathcal{U}_k^\pi := \{(s, a) \in \mathcal{U} \setminus \cup_{i=1}^{k-1} \mathcal{U}_i^\pi : (s, a) \text{ is } \boldsymbol{k_{\text{th}}}\textbf{-certain}\}.$$

We further use $\mathcal{U}_{-(1:K)}^\pi := \mathcal{U} \setminus \cup_{i=1}^{K} \mathcal{U}_i^\pi$ to denote all the remaining state-action pairs in $\mathcal{U}$.

Construct Lookahead Pessimistic MDP

Pessimistic MDP with K-step Lookahead

Definition 3 (LP(ξ, π, K)-MDP) *For any $(\boldsymbol{s}_t, \boldsymbol{a}_t)$, and $\mathcal{M}_{\widehat{p}}$ and $\mathcal{M}_p$, the LP(ξ, π, K)-MDP of $\mathcal{M}_p$, (i.e., $\mathcal{M}^{\pi}_{\widetilde{p}}$) is constructed by modifying the transition $\widehat{p}$ to be $\widetilde{p}$ as:*[a]

<u>**Case 1**</u> *(current transition is* certain*) If $(\boldsymbol{s}_t, \boldsymbol{a}_t) \notin \mathcal{U}$, then*

$$\widetilde{p}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t) = \widehat{p}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t), \tag{9}$$

<u>**Case 2**</u> *(current transition is* uncertain*) If $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}$, then*

<u>**Case 2.1**</u> *(all K-step look ahead is uncertain)*
If $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}^{\pi}_{-(1:K)}$, then

$$\widetilde{p}(\boldsymbol{s}_{t+1} = \mathbf{e}|\boldsymbol{s}_t, \boldsymbol{a}_t) = 1, \tag{10}$$

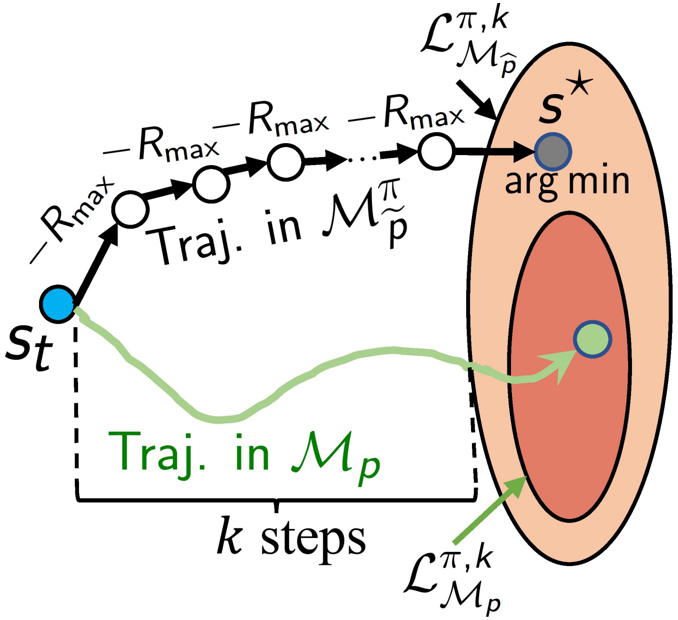
<u>**Case 2.2**</u> *(some k_{th}-step look ahead is certain)*
If $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}^{\pi}_{k}$ for some $k \in [K]$, then we construct a deterministic path such that $\forall\, i \in \{1, ..., k-1\}$,

$$\widetilde{p}(\boldsymbol{s}_{t+k} = \boldsymbol{s}^{\star}|\boldsymbol{s}_t, \boldsymbol{a}_t, \pi) = 1, \;\; \widetilde{r}(\boldsymbol{s}_{t+i}, \cdot) = -R_{\max}. \tag{11}$$

where $\boldsymbol{s}^{\star}$ is defined as:

$$\boldsymbol{s}^{\star} = \underset{\boldsymbol{s}' \in \mathcal{L}^{\pi,k}_{\mathcal{M}_{\widehat{p}}}(\boldsymbol{s}_t, \boldsymbol{a}_t)}{\arg\min} \; V^{\pi}_{\mathcal{M}^{\pi}_{\widetilde{p}}}(\boldsymbol{s}'), \tag{12}$$

[a] The associated reward $\tilde{r}(\cdot) := r(\cdot)$ unless otherwise stated.



Classical Pessimism Principle

- Halts at the uncertain state s_t

Lookahead Pessimism

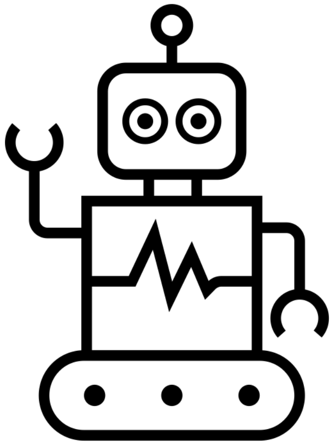
- Constructs a *less conservative path*: $s_t \rightarrow s^\star$
- Has a *better worst-case* guarantee

The Algorithm

Algorithm 1 Policy Learning under Pessimistic MDP with look-ahead (LP(ξ, π, K)-MDP).

Require: Offline data $\mathcal{D}$, threshold ξ, look-ahead steps K, and learning rate α.

1: Fit the dynamics model $\mathcal{M}_{\widehat{p}}$ on $\mathcal{D}$
2: Initialize the policy π_0
3: **for** $n = 0, 1, 2, \ldots$ **do**
4: Construct the LP(ξ, π, K)-MDP for π_n: $\mathcal{M}_{\widetilde{p}}^{\pi_n}$
5: Improve policy: $\pi_{n+1} \leftarrow \texttt{SAC\_Step}(\pi_n, \mathcal{M}_{\widetilde{p}}^{\pi_n})$
6: **end for**
7: **return** π_n



Part 3

Theoretical Analysis

The Suboptimality Bound

Theorem 1 (Performance of the Equilibrium Policy) *Let $\widetilde{\pi}^\star$ denote the equilibrium policy learned under the LP$(\xi, \widetilde{\pi}^\star, K)$-MDP, and let $\pi^\star$ be the optimal policy under the true MDP $\mathcal{M}_p$. Suppose that for any $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}_k^{\widetilde{\pi}^\star}$ with $k \in [1, K]$, $\mathcal{L}_{\mathcal{M}_p}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t) \subseteq \mathcal{L}_{\mathcal{M}_{\widehat{p}}}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ holds, then for any state $\boldsymbol{s}$*

$$
\begin{aligned}
&V_{\mathcal{M}_p}^{\pi^\star}(\boldsymbol{s}) - V_{\mathcal{M}_p}^{\widetilde{\pi}^\star}(\boldsymbol{s}) \\
&\leq \underbrace{\frac{4\gamma\xi R_{\max}}{(1-\gamma)^2}}_{\text{(a)}} + \underbrace{\frac{2\rho^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}} \mathbb{E}\left[\gamma^{\mathcal{T}^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}}}\right] R_{\max}}{1-\gamma}}_{\text{(b) Incurred by hitting } \mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}} \\
&\underbrace{+\sum_{k=1}^{K} \rho^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}} \mathbb{E}\left[\gamma^{\mathcal{T}^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}}\right]\left(\textstyle\sum_{i=1}^{k-1} 2\gamma^i R_{\max} + \gamma^k \Delta_{p,\widetilde{q}}^{\pi^\star,k}(\mathcal{U}_k^{\widetilde{\pi}^\star})\right)}_{\text{(c) Incurred by hitting } \mathcal{U}_k^{\widetilde{\pi}^\star} \text{ for } k\in[K]}.
\end{aligned}
\tag{15}
$$

The Suboptimality Bound

Theorem 1 (Performance of the Equilibrium Policy) *Let $\widetilde{\pi}^\star$ denote the equilibrium policy learned under the LP$(\xi, \widetilde{\pi}^\star, K)$-MDP, and let $\pi^\star$ be the optimal policy under the true MDP $\mathcal{M}_p$. Suppose that for any $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}_k^{\widetilde{\pi}^\star}$ with $k \in [1, K]$, $\mathcal{L}_{\mathcal{M}_p}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t) \subseteq \mathcal{L}_{\mathcal{M}_{\widehat{p}}}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ holds, then for any state $\boldsymbol{s}$*

hitting the **known region** ⋘

$$V_{\mathcal{M}_p}^{\pi^\star}(\boldsymbol{s}) - V_{\mathcal{M}_p}^{\widetilde{\pi}^\star}(\boldsymbol{s})$$

$$\leq \underbrace{\frac{4\gamma\xi R_{\max}}{(1-\gamma)^2}}_{\text{(a)}} + \underbrace{\frac{2\rho^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}} \mathbb{E}\left[\gamma^{\mathcal{T}^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}}}\right] R_{\max}}{1-\gamma}}_{\text{(b) Incurred by hitting } \mathcal{U}^{\widetilde{\pi}^\star}_{-(1:K)}}$$

$$\underbrace{+\sum_{k=1}^{K} \rho^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_k} \mathbb{E}\left[\gamma^{\mathcal{T}^{\pi^\star}_{\boldsymbol{s}\to\mathcal{U}^{\widetilde{\pi}^\star}_k}}\right] \left(\textstyle\sum_{i=1}^{k-1} 2\gamma^i R_{\max} + \gamma^k \Delta^{\pi^\star, k}_{p,\widetilde{q}}(\mathcal{U}^{\widetilde{\pi}^\star}_k)\right)}_{\text{(c) Incurred by hitting } \mathcal{U}^{\widetilde{\pi}^\star}_k \text{ for } k\in[K]}$$

The Suboptimality Bound

Theorem 1 (Performance of the Equilibrium Policy) *Let $\widetilde{\pi}^\star$ denote the equilibrium policy learned under the LP($\xi, \widetilde{\pi}^\star, K$)-MDP, and let $\pi^\star$ be the optimal policy under the true MDP $\mathcal{M}_p$. Suppose that for any $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}_k^{\widetilde{\pi}^\star}$ with $k \in [1, K]$, $\mathcal{L}_{\mathcal{M}_p}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t) \subseteq \mathcal{L}_{\mathcal{M}_{\widehat{p}}}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ holds, then for any state $\boldsymbol{s}$*

$$V_{\mathcal{M}_p}^{\pi^\star}(\boldsymbol{s}) - V_{\mathcal{M}_p}^{\widetilde{\pi}^\star}(\boldsymbol{s})$$

hitting the **known region** ⋘

$$\leq \underbrace{\frac{4\gamma\xi R_{\max}}{(1-\gamma)^2}}_{\text{(a)}} + \underbrace{\frac{2\rho_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star}}\right] R_{\max}}{1-\gamma}}_{\text{(b) Incurred by hitting } \mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}$$

⋙ hitting the **unknown region** & all K-step lookahead is **uncertain**

$$\underbrace{+\sum_{k=1}^{K} \rho_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star}}\right]\left(\sum_{i=1}^{k-1} 2\gamma^i R_{\max} + \gamma^k \Delta_{p,\widetilde{q}}^{\pi^\star, k}(\mathcal{U}_k^{\widetilde{\pi}^\star})\right)}_{\text{(c) Incurred by hitting } \mathcal{U}_k^{\widetilde{\pi}^\star} \text{ for } k\in[K]}$$

The Suboptimality Bound

Theorem 1 (Performance of the Equilibrium Policy) *Let $\widetilde{\pi}^\star$ denote the equilibrium policy learned under the LP$(\xi, \widetilde{\pi}^\star, K)$-MDP, and let $\pi^\star$ be the optimal policy under the true MDP $\mathcal{M}_p$. Suppose that for any $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}_k^{\widetilde{\pi}^\star}$ with $k \in [1, K]$, $\mathcal{L}_{\mathcal{M}_p}^{\widetilde{\pi}^\star,k}(\boldsymbol{s}_t, \boldsymbol{a}_t) \subseteq \mathcal{L}_{\mathcal{M}_{\widehat{p}}}^{\widetilde{\pi}^\star,k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ holds, then for any state $\boldsymbol{s}$*

$$
\begin{aligned}
&V_{\mathcal{M}_p}^{\pi^\star}(\boldsymbol{s}) - V_{\mathcal{M}_p}^{\widetilde{\pi}^\star}(\boldsymbol{s}) \\
&\leq \underbrace{\frac{4\gamma\xi R_{\max}}{(1-\gamma)^2}}_{\text{(a)}} + \underbrace{\frac{2\rho_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star}}\right] R_{\max}}{1-\gamma}}_{\text{(b) Incurred by hitting } \mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}} \\
&+ \underbrace{\sum_{k=1}^{K} \rho_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star}}\right]\left(\sum_{i=1}^{k-1} 2\gamma^i R_{\max} + \gamma^k \Delta_{p,\widetilde{q}}^{\pi^\star,k}(\mathcal{U}_k^{\widetilde{\pi}^\star})\right)}_{\text{(c) Incurred by hitting } \mathcal{U}_k^{\widetilde{\pi}^\star} \text{ for } k\in[K]}
\end{aligned}
$$

hitting the **known region** <<< (a)

>>> hitting the **unknown region** & all K-step lookahead is **uncertain** (b)

>>> hitting the **unknown region** & some k-step lookahead is **certain** (c)

The Suboptimality Bound

Theorem 1 (Performance of the Equilibrium Policy) *Let $\widetilde{\pi}^\star$ denote the equilibrium policy learned under the LP$(\xi, \widetilde{\pi}^\star, K)$-MDP, and let $\pi^\star$ be the optimal policy under the true MDP $\mathcal{M}_p$. Suppose that for any $(\boldsymbol{s}_t, \boldsymbol{a}_t) \in \mathcal{U}_k^{\widetilde{\pi}^\star}$ with $k \in [1, K]$, $\mathcal{L}_{\mathcal{M}_p}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t) \subseteq \mathcal{L}_{\mathcal{M}_{\widehat{p}}}^{\widetilde{\pi}^\star, k}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ holds, then for any state $\boldsymbol{s}$*

$$V_{\mathcal{M}_p}^{\pi^\star}(\boldsymbol{s}) - V_{\mathcal{M}_p}^{\widetilde{\pi}^\star}(\boldsymbol{s})$$

$$\leq \underbrace{\frac{4\gamma\xi R_{\max}}{(1-\gamma)^2}}_{\text{(a)}} + \underbrace{\frac{2\rho_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}^{\pi^\star}}\right] R_{\max}}{1-\gamma}}_{\text{(b) Incurred by hitting } \mathcal{U}_{-(1:K)}^{\widetilde{\pi}^\star}}$$

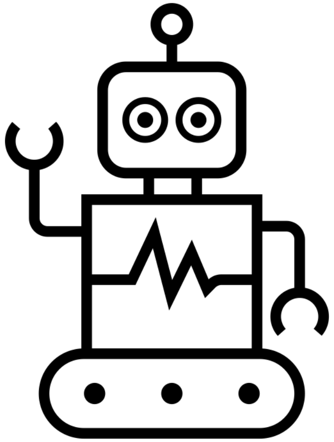
$$+ \underbrace{\sum_{k=1}^{K} \rho_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star} \mathbb{E}\left[\gamma^{\mathcal{T}_{\boldsymbol{s}\to\mathcal{U}_k^{\widetilde{\pi}^\star}}^{\pi^\star}}\right] \left(\sum_{i=1}^{k-1} 2\gamma^i R_{\max} + \gamma^k \Delta_{p,\widetilde{q}}^{\pi^\star,k}(\mathcal{U}_k^{\widetilde{\pi}^\star})\right)}_{\text{(c) Incurred by hitting } \mathcal{U}_k^{\widetilde{\pi}^\star} \text{ for } k\in[K]}$$

hitting the **known region** ⋘ (a)

(b) ⋙ hitting the **unknown region** & all K-step lookahead is **uncertain**

(c) ⋙ hitting the **unknown region** & some k-step lookahead is **certain**

💡 Monotonically improves with K → guaranteed ***better lower bound*** than existing work!

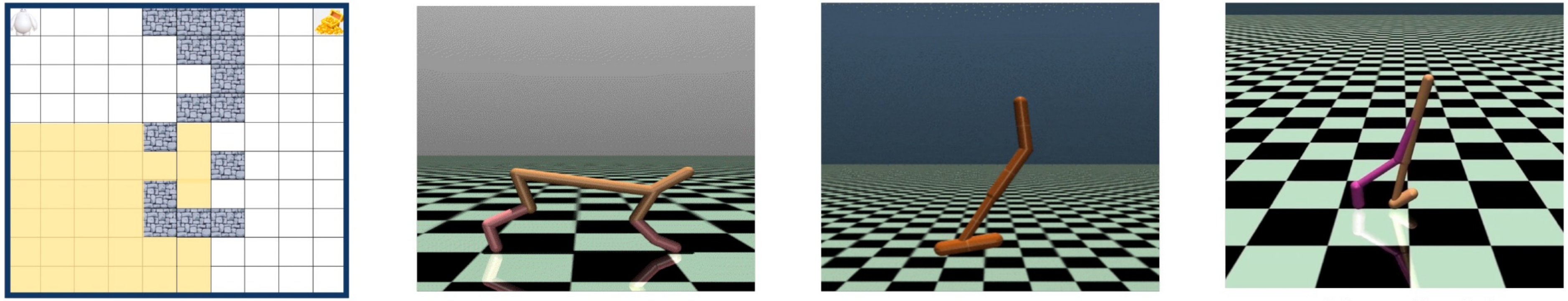


Part 4

Experimental Results

Datasets

Popular benchmark suite used in many papers (Kidambi et al., 2020, etc.).



(a) Grid world (b) Halfcheetah (c) Hopper (d) Walker2d

Figure 4. Visualization of considered tasks. (a) The grid world task is adapted from Eysenbach et al. (2022) with increased difficulty. The agent starts where the baymax is located (top left corner). The reward for hitting the treasure is +50, +1 for yellow-shaded grids, and +0.5 for other grids. The walls cannot be crossed. For (b), (c) and (d), the tasks come from the D4RL benchmark (Fu et al., 2020).

Results: Performance Improvement

Dataset	Environment	LP-MDP (Ours)	MOReL (SAC)	MOReL (NPG)	MOPO	CQL	SAC-Off	BEAR	BRAC-p	BRAC-v
medium	halfcheetah	42.6±5.5	43.4	42.1	54.2	44.4	-4.3	41.7	43.8	46.3
medium	hopper	101.9 ±1.1*	75.8	95.4	28.0	86.6	0.8	52.1	32.7	31.1
medium	walker2d	64.5±5.1	76.8	77.8	17.8	74.5	0.9	59.1	77.5	81.1
medium-replay	halfcheetah	48.5±2.1*	43.4	40.2	53.1	46.2	-2.4	38.6	45.4	47.7
medium-replay	hopper	101.2 ±0.8	101.1	93.6	67.5	48.6	3.5	33.7	0.6	0.6
medium-replay	walker2d	82.7 ±5.9*	46.5	49.8	39.0	32.6	1.9	19.2	-0.3	0.9
medium-expert	halfcheetah	51.1±0.9*	41.6	53.3	63.3	62.4	1.8	53.4	44.2	41.9
medium-expert	hopper	103.7±1.2*	78.5	108.7	23.7	111.0	1.6	96.3	1.9	0.8
medium-expert	walker2d	81.5±8.5*	68.0	95.6	44.6	98.7	-0.1	40.1	76.9	81.6
Average Scores		677.7 *	575.1	656.5	391.2	605.0	3.7	434.2	328.1	332.0

Table 1. Results on D4RL. We report the averaged normalized scores of 3 different random seeds with one standard errors. We highlight the best averaged scores by a blue box . Also, we use * to indicate the tasks where our method has a solid improvement over MOReL-SAC.

Results: The Effect of Lookahead Horizon K

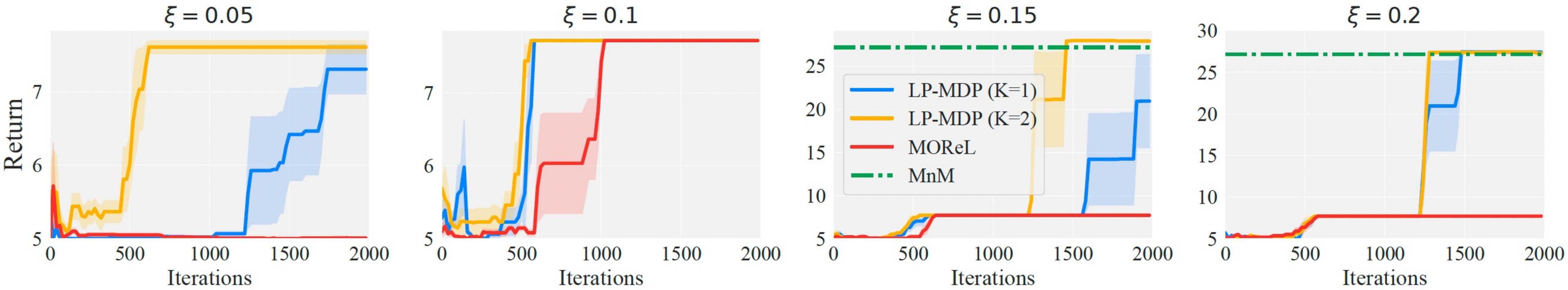


Figure 5. Results on the gridworld task. The dashed green curve is the performance of the expert policy obtained with MnM (Eysenbach et al., 2022). The shaded region is the one standard error of three trials with different random seeds.

Policy Behaviors

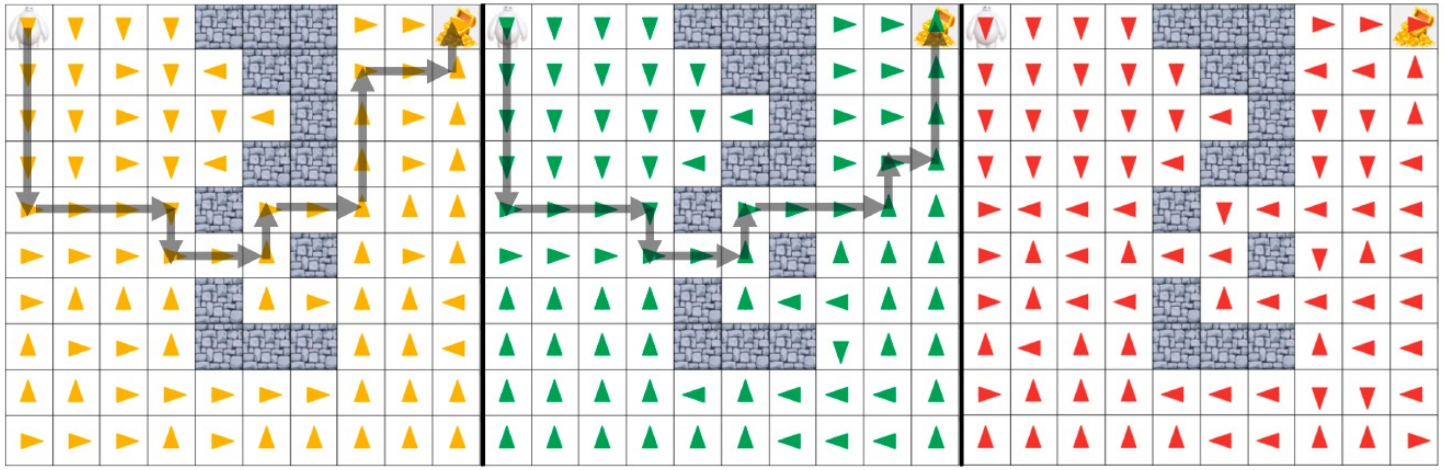


Figure 6. A visualization of the policies. The yellow policy (left) is the one learnt under LP-MDP with $k = 2$ and $\xi = 0.15$. In the middle, the green policy is the expert policy learnt using MnM (Eysenbach et al., 2022). Lastly, the red policy corresponds to the one learnt using MOReL. The grey trajectories are possible paths from the starting grid to the treasure by following the learned policies.

Summary

Key Takeaways

- Don’t be pessimistic too early; be far-sighted!
- Classical pessimism principle can result in rather sub-optimal policy.
- Model lookahead helps modulate pessimism, giving better guarantee.

Future directions

- Offline-to-Online RL.
- RL as sequence modeling problem (*i.e.*, return/goal conditioned imitation learning).

A Provably Better Offline RL Algorithm

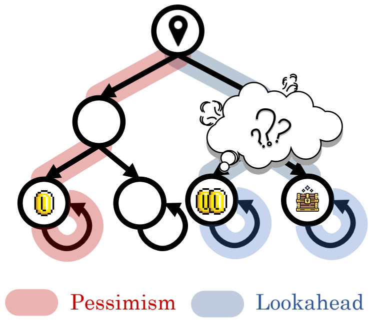
Don't be Pessimistic Too Early: Look K Steps Ahead!

Chaoqi Wang
University of Chicago

Ziyu Ye
University of Chicago

Kevin Murphy
Google Research

Yuxin Chen
University of Chicago



February 2023