Model-based RL as a Minimalist Approach to Horizon-Free and Second-Order Bounds

Zhiyong Wang[1], Dongruo Zhou[2], John C.S. Lui[1], Wen Sun[3]

[1]The Chinese University of Hong Kong (CUHK) [2]Indiana University
[3]Cornell University

ICLR 2025

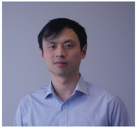
Selected as a course reference paper for CS 6789:
Foundations of Reinforcement Learning at Cornell University

May, 2025

Zhiyong Wang

Dongruo Zhou

John C.S. Lui



Wen Sun

Outline

- Preliminaries
- Online RL
- Offline RL
- Proof Sketch
- Summary

Horizon-free and Second-order MBRL: Preliminaries

- We consider finite horizon time-homogenous MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, P^\star, r, s_0\}$
  - $\mathcal{S}, \mathcal{A}$ are the state and action space
  - $H \in \mathbb{N}^+$ is the horizon for each episode
  - $P^\star : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the ground truth unknown transition
  - $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the known reward signal, and s_0 is the fixed initial state.
- At each episode, the agent interacts with the environment over a sequence of H time steps. Specifically, starting from the initial state s_0, at each time step $h \in [H-1]$,
  - the agent observes the current state $s_h \in \mathcal{S}$,
  - takes an action $a_h = \pi_h(s_h) \in \mathcal{A}$ according to its policy,
  - receives a reward $r(s_h, a_h)$, and
  - the environment transitions to the next state $s_{h+1} \sim P^\star(\cdot \mid s_h, a_h)$.
  - The cumulative reward over the episode is defined as $\sum_{h=0}^{H-1} r(s_h, a_h)$.

Horizon-free and Second-order MBRL: Preliminaries

- $V_h^\pi(s)$ represents the expected total reward of policy π starting at $s_h = s$
- $Q_h^\pi(s,a)$ is the expected total reward of the process of executing a at s at time step h followed by executing π to the end.
- The optimal policy $\pi^\star$ is defined as $\pi^\star = \mathrm{argmax}_\pi V_0^\pi(s_0)$.
- Since we use the model-based approach for learning, we define a general model class $\mathcal{P} \subset \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$.
- Given a transition P, we denote $V_{h;P}^\pi$ and $Q_{h;P}^\pi$ as the value and Q functions of policy π under the model P.
- Given a function $f : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, we denote the $(Pf)(s,a) := \mathbb{E}_{s' \sim P(s,a)} f(s')$. We then denote the variance induced by one-step transition P and function f as $(\mathbb{V}_P f)(s,a) := \left(Pf^2\right)(s,a) - \left(Pf(s,a)\right)^2$ which is equal to $\mathbb{E}_{s' \sim P(s,a)} f^2(s') - \left(\mathbb{E}_{s' \sim P(s,a)} f(s')\right)^2$.

Horizon-free and Second-order MBRL: Preliminaries

- Assumptions:
  1. Realizability: $P^\star \in \mathcal{P}$.
  2. We assume that the rewards are normalized such that $r(\tau) \in [0,1]$ for any trajectory $\tau := \{s_0, a_0, \ldots, s_{H-1}, a_{H-1}\}$ where $r(\tau)$ is short for $\sum_{h=0}^{H-1} r(s_h, a_h)$.

Horizon-free and Second-order MBRL: Preliminaries

- Online RL:
  1. We focus on the episodic setting where the learner can interact with the environment for K episodes. At episode k, the learner proposes a policy π^k (based on the past interaction history), executes π^k starting from s_0 to time step $H-1$.
  2. We measure the performance of the online learning via regret: $\sum_{k=0}^{K-1}\left(V^{\pi^\star} - V^{\pi^k}\right)$.
  3. To achieve meaningful regret bounds, we often need additional structural assumptions on the MDP and the model class $\mathcal{P}$. We use the ℓ_1 Eluder dimension as the structural condition [1].

Horizon-free and Second-order MBRL: Preliminaries

- Offline RL:
  1. For the offline RL setting, we assume that we have a pre-collected offline dataset $\mathcal{D} = \{\tau^i\}_{i=1}^K$ which contains K trajectories.
  2. To succeed in offline learning, we typically require the offline dataset to have good coverage over some high-quality comparator policy π^*.
  3. Our goal is to learn a policy $\widehat{\pi}$ that is as good as π^*, and we are interested in the performance gap between $\hat{\pi}$ and π^*, i.e., $V^{\pi^*} - V^{\hat{\pi}}$.

Horizon-free and Second-order MBRL: Preliminaries

- Horizon-free Bound:
  1. The regret or sample complexity bounds have no explicit polynomial dependence on the horizon H.
  2. Motivation: to see if RL problems are harder than contextual bandits due to the longer horizon planning in RL.
  3. Some previous works use extremely complex algorithms and analysis in the tabular MDP case[1].
- Second-order Bound:
  1. Denote VaR_π as the variance of trajectory reward, i.e., $\mathrm{VaR}_\pi := \mathbb{E}_{\tau\sim\pi}(r(\tau) - \mathbb{E}_{\tau\sim\pi} r(\tau))^2$. Second-order bounds in offline RL scales with VaR_{π^*} – the variance of the comparator policy. Second-order regret bound in online setting scales with respect to $\sqrt{\sum_k \mathrm{VaR}_{\pi^k}}$ instead of $\sqrt{K}$.
  2. The second-order bound can be small under situations such as nearly-deterministic systems or the optimal policy having a small value.

[1] e.g., [1] Settling the Horizon-Dependence of Sample Complexity in Reinforcement Learning, FOCS 2021
and [2] Horizon-Free Reinforcement Learning in Polynomial Time: the Power of Stationary Policies, COLT 2022

The key message of our work is

Simple and standard MLE-based MBRL algorithms are sufficient for achieving nearly horizon-free and second-order bounds in online and offline RL with function approximation.

Horizon-free and Second-order MBRL: Online Setting

- At episode k, O-MBRL splits the trajectory data that contains $k-1$ trajectories into a dataset of (s, a, s') tuples which is used to perform maximum likelihood estimation $\max_{\tilde{P} \in \mathcal{P}} \sum_{i=1}^{n} \log \tilde{P}(s_i'|s_i, a_i)$.
- It then builds a version space $\widehat{\mathcal{P}}^k$ which contains models $P \in \mathcal{P}$ whose log data likelihood is not below by too much than that of the MLE estimator.
- The version space is designed such that for all $k \in [0, K-1]$, we have $P^\star \in \widehat{\mathcal{P}}_k$ with high probability.
- The policy π^k in this case is computed via the optimism principle.

Algorithm 1 Optimistic Model-based RL (O-MBRL)

1: **Input:** model class $\mathcal{P}$, confidence parameter $\delta \in (0,1)$, threshold β.
2: Initialize π^0, initialize dataset $\mathcal{D} = \varnothing$.
3: **for** $k = 0 \to K-1$ **do**
4: Collect a trajectory $\tau = \{s_0, a_0, \cdots, s_{H-1}, a_{H-1}\}$ from π^k, split it into tuples of $\{s, a, s'\}$ and add to $\mathcal{D}$.
5: Construct a version space $\widehat{\mathcal{P}}^k$:

$$\widehat{\mathcal{P}}^k = \left\{ P \in \mathcal{P} : \sum_{s,a,s' \in \mathcal{D}} \log P(s_i'|s_i, a_i) \geq \max_{\tilde{P} \in \mathcal{P}} \sum_{s,a,s' \in \mathcal{D}} \log \tilde{P}(s_i'|s_i, a_i) - \beta \right\}.$$

6: Set $(\pi^k, \widehat{P}^k) \leftarrow \operatorname{argmax}_{\pi \in \Pi, P \in \widehat{\mathcal{P}}^k} V^{\pi}_{0;P}(s_0)$.
7: **end for**

Horizon-free and Second-order MBRL: Online Setting

- We work with the ℓ_1 Eluder dimension $DE_1(\Psi, \mathcal{S} \times \mathcal{A}, \epsilon)$ with the function class Ψ specified as:

$$\Psi = \left\{(s, a) \mapsto \mathbb{H}^2(P^\star(s, a) \parallel P(s, a)) : P \in \mathcal{P}\right\}.$$

Remark

The ℓ_1 Eluder dimension has been widely used in previous works [1]. It can capture tabular, linear, and generalized linear models.

Horizon-free and Second-order MBRL: Online Setting

Theorem (Main theorem for online setting)

For any $\delta \in (0,1)$, let $\beta = 4\log\left(\frac{K|\mathcal{P}|}{\delta}\right)$, with probability at least $1-\delta$, O-MBRL achieves the following regret bound:

$$\sum_{k=0}^{K-1}(V^{\pi^\star} - V^{\pi^k}) \leq O\Big(\sqrt{\sum_{k=0}^{K-1} \mathrm{VaR}_{\pi^k} \cdot \mathrm{DE}_1(\Psi, \mathcal{S}\times\mathcal{A}, 1/KH) \cdot \log(KH\,|\mathcal{P}|\,/\delta)\log(KH)}$$
$$+ \mathrm{DE}_1(\Psi, \mathcal{S}\times\mathcal{A}, 1/KH) \cdot \log(KH\,|\mathcal{P}|\,/\delta)\log(KH)\Big). \tag{1}$$

- The above theorem indicates the standard and simple O-MBRL algorithm is already enough to achieve horizon-free and second-order regret bounds: our bound does not have explicit polynomial dependences on horizon H, the leading term scales with $\sqrt{\sum_k \mathrm{VaR}_{\pi^k}}$ instead of the typical $\sqrt{K}$.

Horizon-free and Second-order MBRL: Online Setting

When the underlying MDP has deterministic transitions, we can achieve a smaller regret bound that only depends on the number of episodes logarithmically.

Corollary (log K regret bound with deterministic transitions)

When the transition dynamics of the MDP are deterministic, setting $\beta = 4\log\left(\frac{K|\mathcal{P}|}{\delta}\right)$, w.p. at least $1-\delta$, O-MBRL achieves:

$$\sum_{k=0}^{K-1} V^{\pi^\star} - V^{\pi^k} \le O\left(\mathrm{DE}_1(\Psi, \mathcal{S}\times\mathcal{A}, 1/KH)\cdot \log(KH\left|\mathcal{P}\right|/\delta)\log(KH)\right).$$

Horizon-free and Second-order MBRL: Offline Setting

- CPPO-LR splits the offline trajectory data that contains K trajectories into a dataset of (s, a, s') tuples which is used to perform maximum likelihood estimation $\max_{\tilde{P}\in\mathcal{P}} \sum_{i=1}^{n} \log \tilde{P}(s'_i|s_i, a_i)$.
- It then builds a version space $\widehat{\mathcal{P}}$ which contains models $P \in \mathcal{P}$ whose log data likelihood is not below by too much than that of the MLE estimator.
- The threshold for the version space is constructed so that with high probability, $P^\star \in \widehat{\mathcal{P}}$.
- Once we build a version space, we perform pessimistic planning to compute $\widehat{\pi}$.

Algorithm 2 (Uehara & Sun (2021)) Constrained Pessimistic Policy Optimization with Likelihood-Ratio based constraints (CPPO-LR)

1: **Input:** dataset $\mathcal{D} = \{s, a, s'\}$, model class $\mathcal{P}$, policy class Π, confidence parameter $\delta \in (0, 1)$, threshold β.

2: Calculate the confidence set based on the offline dataset:

$$\widehat{\mathcal{P}} = \left\{ P \in \mathcal{P} : \sum_{i=1}^{n} \log P(s'_i|s_i, a_i) \geq \max_{\tilde{P}\in\mathcal{P}} \sum_{i=1}^{n} \log \tilde{P}(s'_i|s_i, a_i) - \beta \right\}.$$

3: **Output:** $\hat{\pi} \leftarrow \mathrm{argmax}_{\pi\in\Pi} \min_{P\in\widehat{\mathcal{P}}} V^{\pi}_{0;P}(s_0)$.

Horizon-free and Second-order MBRL: Offline Setting

Definition (Single policy coverage)

Given any comparator policy π^*, denote the data-dependent single policy concentrability coefficient $C_{\mathcal{D}}^{\pi^*}$ as follows:

$$C_{\mathcal{D}}^{\pi^*} := \max_{h, P \in \mathcal{P}} \frac{\mathbb{E}_{s,a \sim d_h^{\pi^*}} \mathbb{H}^2 \left(P(s,a) \parallel P^\star(s,a)\right)}{1/K \sum_{k=1}^{K} \mathbb{H}^2 \left(P(s_h^k, a_h^k) \parallel P^\star(s_h^k, a_h^k)\right)}.$$

Theorem (Performance gap of CPPO-LR)

For any $\delta \in (0,1)$, let $\beta = 4 \log(|\mathcal{P}|/\delta)$, w.p. at least $1-\delta$, CPPO-LR learns a policy $\widehat{\pi}$ that enjoys the following performance gap with respect to any comparator policy π^*:

$$V^{\pi^*} - V^{\widehat{\pi}} \leq O\left(\sqrt{C^{\pi^*} \mathrm{VaR}_{\pi^*} \log(|\mathcal{P}|/\delta)/K} + C^{\pi^*} \log(|\mathcal{P}|/\delta)/K\right).$$

Horizon-free and Second-order MBRL: Offline Setting

- First, our bound is horizon-free (not even any $\log(H)$ dependence), while the previous bound in [2] has $\text{poly}(H)$ dependence.
- Second, our bound scales with $\text{VaR}_{\pi^*} \in [0, 1]$, which can be small when $\text{VaR}_{\pi^*} \ll 1$.

Corollary (C^{π^*}/K performance gap of CPPO-LR with deterministic transitions)

When the ground truth transition $P^\star$ of the MDP is deterministic, for any $\delta \in (0, 1)$, let $\beta = 4 \log(|\mathcal{P}|/\delta)$, w.p. at least $1 - \delta$, CPPO-LR learns a policy $\widehat{\pi}$ that enjoys the following performance gap with respect to any comparator policy π^*:

$$V^{\pi^*} - V^{\widehat{\pi}} \leq O\left(C^{\pi^*} \log(|\mathcal{P}|/\delta)/K\right).$$

Proof Sketch for Online RL

- For ease of presentation, we use d_{RL} to denote $\mathrm{DE}_1(\Psi, \mathcal{S} \times \mathcal{A}, 1/KH)$, and ignore some log terms.
- Overall, our analysis follows the general framework of optimism in the face of uncertainty, but with
  1. careful analysis in leveraging the MLE generalization bound
  2. novel analyses to achieve a variance-dependent bound without estimating variances
  3. a more refined proof in the training-to-testing distribution transfer via Eluder dimension
  4. careful variance recursion analysis.

Proof Sketch for Online RL

- By standard MLE analysis, we can show w.p. $1-\delta$, for all $k \in [K-1]$, we have $P^\star \in \widehat{\mathcal{P}}^k$, and

$$\sum_{i=0}^{k-1} \sum_{h=0}^{H-1} \mathbb{H}^2(P^\star(s_h^i, a_h^i) || \widehat{P}^k(s_h^i, a_h^i)) \leq O(\log(K \left|\mathcal{P}\right| / \delta)) \,. \tag{2}$$

- From here, trivially applying training-to-testing distribution transfer via the Eluder dimension as previous works would cause poly-dependence on H.
- With some new techniques, we can get: there exists a set $\mathcal{K} \subseteq [K-1]$ such that $|\mathcal{K}| \leq O(d_{\mathrm{RL}} \log(K|\mathcal{P}|/\delta))$, and

$$\sum_{k \in [K-1] \setminus \mathcal{K}} \sum_{h} \mathbb{H}^2\Big(P^\star(s_h^k, a_h^k) \parallel \widehat{P}^k\big(s_h^k, a_h^k\big)\Big) \leq O(d_{\mathrm{RL}} \cdot \log(K \left|\mathcal{P}\right| / \delta) \log(KH)) \,. \tag{3}$$

Proof Sketch for Online RL

- Recall that $(\pi^k, \widehat{P}^k) \leftarrow \mathrm{argmax}_{\pi\in\Pi, P\in\widehat{\mathcal{P}}^k} V^{\pi}_{0;P}(s_0)$, with the realization guarantee $P^\star \in \widehat{\mathcal{P}}^k$, we can get the following optimism guarantee: $V^\star_{0;P^\star} \leq \max_{\pi\in\Pi, P\in\widehat{\mathcal{P}}^k} V^{\pi}_{0;P} = V^{\pi^k}_{0;\widehat{P}^k}$.
- At this stage, one straight-forward way to proceed is to use the standard simulation lemma:

$$\begin{aligned}
&\sum_{k=0}^{K-1} V^{\pi^\star}_{0;P^\star} - V^{\pi^k}_{0;P^\star} \leq \sum_{k=0}^{K-1} V^{\pi^k}_{0;\widehat{P}^k} - V^{\pi^k}_{0;P^\star} \\
&\leq \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim d_h^{\pi^k}}\left[\left|\mathbb{E}_{s'\sim P^\star(s,a)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s'\sim \widehat{P}^k(s,a)} V^{\pi^k}_{h+1;\widehat{P}^k}(s')\right|\right]. \qquad (4)
\end{aligned}$$

- However, from here, if we naively bound each term on the RHS via $\mathbb{E}_{s,a\sim d_h^{\pi^k}}\|P^\star(s,a) - \widehat{P}^k(s,a)\|_1$, which is what previous works such as [2] did exactly, we would end up paying a polynomial horizon dependence H due to the summation over H on the RHS the above expression.

Proof Sketch for Online RL

- We have the following mean-to-variance lemma

Lemma (Lemma 4.3 in [3])

For two distributions $f \in \Delta([0,1])$ and $g \in \Delta([0,1])$:

$$|\mathbb{E}_{x\sim f}[x] - \mathbb{E}_{x\sim g}[x]| \leq 4\sqrt{\mathrm{VaR}_f \cdot D_\triangle(f \parallel g)} + 5D_\triangle(f \parallel g). \tag{5}$$

where $\mathrm{VaR}_f := \mathbb{E}_{x\sim f}(x - \mathbb{E}_{x\sim f}[x])^2$ denotes the variance of the distribution f.

- Given this mean-to-variance lemma, we may consider using it to bound the difference between two means $\mathbb{E}_{s'\sim P^\star(s,a)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s'\sim \widehat{P}^k(s,a)} V^{\pi^k}_{h+1;\widehat{P}^k}(s')$.
- This still can not work if we start from here, because we would eventually get $\sum_k \sum_h \mathbb{E}_{s,a\sim d_h^{\pi^k}}[\mathbb{H}^2(P^\star(s,a)||\widehat{P}^k(s,a))]$ terms, which can not be further upper bounded easily with the MLE generalization guarantee.

Proof Sketch for Online RL

- To achieve horizon-free and second-order bounds, we need a novel and more careful analysis.
- First, we carefully decompose and upper bound the regret in $\tilde{\mathcal{K}} := [K-1] \setminus \mathcal{K}$ w.h.p. as follows using Bernstein's inequality (for regret in $\mathcal{K}$ we can simply upper bound it by $|\mathcal{K}|$)

$$
\begin{aligned}
&\sum_{k \in \tilde{\mathcal{K}}} \left(V^{\pi^k}_{0;\widehat{P}^k}(s_h^k) - \sum_{h=0}^{H-1} r(s_h^k, a_h^k) \right) + \sum_{k \in \tilde{\mathcal{K}}} \left(\sum_{h=0}^{H-1} r(s_h^k, a_h^k) - V^{\pi^k}_{0;P^\star} \right) \\
&\lesssim \sqrt{\sum_{k \in \tilde{\mathcal{K}}} \sum_h \left(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1;\widehat{P}^k} \right)(s_h^k, a_h^k)} \\
&+ \sum_{k \in \tilde{\mathcal{K}}} \sum_h \left| \mathbb{E}_{s' \sim \widehat{P}^k(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s' \sim P^*(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') \right| \\
&+ \sqrt{\sum_k \mathrm{VaR}_{\pi^k} \log(1/\delta)}\,.
\end{aligned}
\tag{6}
$$

Proof Sketch for Online RL

- Then, we bound the difference of two means $\mathbb{E}_{s' \sim \widehat{P}^k(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s' \sim P^*(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s')$ using variances and the triangle discrimination using the mean-to-variance lemma, together with the fact that $D_\triangle \le 4\mathbb{H}^2$, and information processing inequality on the squared Hellinger distance, we have

$$
\begin{aligned}
&|\mathbb{E}_{s' \sim \widehat{P}^k(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s' \sim P^*(s_h^k, a_h^k)} V^{\pi^k}_{h+1;\widehat{P}^k}(s')| \\
&\le O\Big(\sqrt{(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1;\widehat{P}^k})(s_h^k, a_h^k) D_\triangle\Big(V^{\pi^k}_{h+1;\widehat{P}^k}(s' \sim P^\star(s_h^k, a_h^k)) \parallel V^{\pi^k}_{h+1;\widehat{P}^k}(s' \sim \widehat{P}^k(s_h^k, a_h^k))\Big)} \\
&\qquad + D_\triangle\Big(V^{\pi^k}_{h+1;\widehat{P}^k}(s' \sim P^\star(s_h^k, a_h^k)) \parallel V^{\pi^k}_{h+1;\widehat{P}^k}(s' \sim \widehat{P}^k(s_h^k, a_h^k))\Big)\Big) \\
&\le O\Big(\sqrt{(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1;\widehat{P}^k})(s_h^k, a_h^k) \mathbb{H}^2\Big(P^\star(s_h^k, a_h^k) \parallel \widehat{P}^k(s_h^k, a_h^k)\Big)} + \mathbb{H}^2\Big(P^\star(s_h^k, a_h^k) \parallel \widehat{P}^k(s_h^k, a_h^k)\Big)\Big)
\end{aligned}
$$

where we denote $V^{\pi^*}_{h+1;\widehat{P}}(s' \sim P^\star(s,a))$ as the distribution of the random variable $V^{\pi^*}_{h+1;\widehat{P}}(s')$ with $s' \sim P^\star(s,a)$.

Proof Sketch for Online RL

- Then, summing up over k, h, with Cauchy-Schwartz and the MLE generalization bound via Eluder dimension in Eq.(3), we have

$$\begin{aligned}
&\sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\left|\mathbb{E}_{s'\sim\widehat{P}^k(s_h^k,a_h^k)}V^{\pi^k}_{h+1;\widehat{P}^k}(s') - \mathbb{E}_{s'\sim P^*(s_h^k,a_h^k)}V^{\pi^k}_{h+1;\widehat{P}^k}(s')\right| \\
&\le O\Bigg(\sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\mathbb{H}^2\Big(P^\star(s_h^k,a_h^k)\parallel\widehat{P}^k\big(s_h^k,a_h^k\big)\Big) \\
&+\sqrt{\sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\big(\mathbb{V}_{P^\star}V^{\pi^k}_{h+1;\widehat{P}^k}\big)(s_h^k,a_h^k)\sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\mathbb{H}^2\Big(P^\star(s_h^k,a_h^k)\parallel\widehat{P}^k\big(s_h^k,a_h^k\big)\Big)}\Bigg) \\
&\le O\Bigg(\sqrt{\sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\big(\mathbb{V}_{P^\star}V^{\pi^k}_{h+1;\widehat{P}^k}\big)(s_h^k,a_h^k)d_{\mathrm{RL}}\log(K\left|\mathcal{P}\right|/\delta)\log(KH)} \\
&+ d_{\mathrm{RL}}\log(K\left|\mathcal{P}\right|/\delta)\log(KH)\Bigg). \qquad (7)
\end{aligned}$$

Proof Sketch for Online RL

- Note that we have $\big(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1;\widehat{P}^k}\big)(s^k_h, a^k_h)$ depending on $\widehat{P}^k$. To get a second-order bound, we need to convert it to the variance under ground truth transition $P^\star$, and we want to do it without incurring any H dependence.
- We aim to replace $\big(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1;\widehat{P}^k}\big)(s^k_h, a^k_h)$ by $\big(\mathbb{V}_{P^\star} V^{\pi^k}_{h+1}\big)(s^k_h, a^k_h)$ which is the variance under $P^\star$, and we want to control the difference $\Big(\mathbb{V}_{P^\star}\Big(V^{\pi^k}_{h+1;\widehat{P}^k} - V^{\pi^k}_{h+1}\Big)\Big)(s^k_h, a^k_h)$. To do so, we need to bound the variance of the 2^m-th moment of the difference $V^{\pi^k}_{h+1;\widehat{P}^k} - V^{\pi^k}_{h+1}$.

Proof Sketch for Online RL

- Let us define the following terms:

$$A := \sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\left[\left(\mathbb{V}_{P^\star}V^{\pi^k}_{h+1;\widehat{P}^k}\right)(s^k_h,a^k_h)\right], B := \sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\left[\left(\mathbb{V}_{P^\star}V^{\pi^k}_{h+1}\right)(s^k_h,a^k_h)\right],$$

$$C_m := \sum_{k\in\tilde{\mathcal{K}}}\sum_{h}\left[\left(\mathbb{V}_{P^\star}(V^{\pi^k}_{h+1;\widehat{P}^k} - V^{\pi^k}_{h+1})^{2^m}\right)(s^k_h,a^k_h)\right],$$

$$G := \sqrt{A\cdot d_{\mathrm{RL}}\log(\frac{K\,|\mathcal{P}|}{\delta})\log(KH)} + d_{\mathrm{RL}}\log(\frac{K\,|\mathcal{P}|}{\delta})\log(KH)\,.$$

- With the fact $\mathbb{V}_{P^\star}(a+b) \le 2\mathbb{V}_{P^\star}(a) + 2\mathbb{V}_{P^\star}(b)$ we have $A \le 2B + 2C_0$.

Proof Sketch for Online RL

- For C_m, we prove that w.h.p. it has the recursive form $C_m \lesssim 2^m G + \sqrt{\log(1/\delta) C_{m+1}} + \log(1/\delta)$, during which process we also leverage the above Eq.(7) and some careful analysis.
- Then, with a recursion lemma, we can get $C_0 \lesssim G$, which further gives us

$$
\begin{aligned}
A &\lesssim B + d_{\mathrm{RL}} \log\left(\frac{K\,|\mathcal{P}|}{\delta}\right) \log(KH) + \sqrt{A \cdot d_{\mathrm{RL}} \log\left(\frac{K\,|\mathcal{P}|}{\delta}\right) \log(KH)} \\
&\leq O\big(B + d_{\mathrm{RL}} \log\left(\frac{K\,|\mathcal{P}|}{\delta}\right) \log(KH)\big) ,
\end{aligned}
$$

  where in the last step we use the fact $x \leq 2a + b^2$ if $x \leq a + b\sqrt{x}$.
- Finally, we note that $B \leq O(\sum_k \mathrm{VaR}_{\pi^k} + \log(1/\delta))$ w.h.p.. Plugging the upper bound of A back into Eq.(7) and then to Eq.(6), we conclude the proof.

Overall, our work identifies the minimalist algorithms and analysis for nearly horizon-free and second-order online & offline RL.

Horizon-free and Second-order MBRL: Summary

- There are some interesting future works:
  1. Remove the $\log H$ dependence (completely horizon-free).
  2. Extend our analysis to incorporate the richer function classes with small distributional Eduler dimensions.
  3. The algorithms studied in this work are not computationally tractable. This is due to the need of performing optimism/pessimism planning. Deriving computationally tractable RL algorithms for the rich function approximation setting is a long-standing question.

Thank you!

References

[1] Qinghua Liu et al. “When is partially observable reinforcement learning not scary?” In: Conference on Learning Theory. PMLR. 2022, pp. 5175–5220.

[2] Masatoshi Uehara and Wen Sun. “Pessimistic model-based offline reinforcement learning under partial coverage”. In: arXiv preprint arXiv:2107.06226 (2021).

[3] Kaiwen Wang et al. “More Benefits of Being Distributional: Second-Order Bounds for Reinforcement Learning”. In: arXiv preprint arXiv:2402.07198 (2024).