Towards better Instance-dependent offline Reinforcement Learning

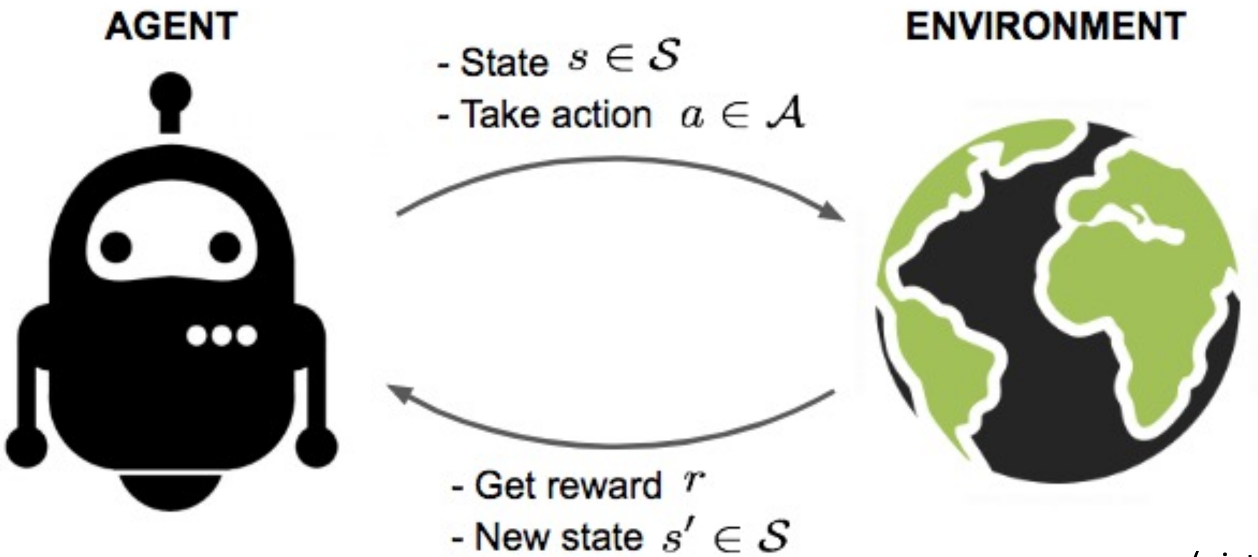
Ming Yin

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Outline

1. Overview of our works
2. Preliminary
3. Tabular Setting
4. Linear MDP setting
5. Summary

Reinforcement Learning



(picture from internet)

An RL agent learns interactively through the feedbacks of an environment.

And in real-life applications as well…

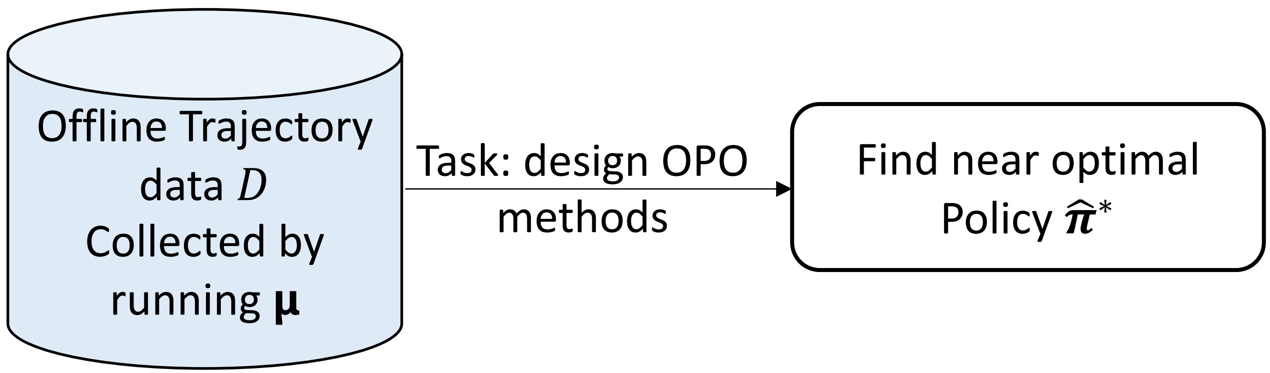
- RL for robotics.
- RL for dialogue systems.
- RL for personalized medicine.
- RL for self-driving cars.
- RL for new material discovery.
- RL for sustainable energy.
- RL for feature-based dynamic pricing.
- RL for maximizing user satisfaction.
- RL for QoE optimization in networking
- …

However, there are Challenges…

- No access to a simulator
- Every data point is costly.
- Legal, safety issues associated with exploration.
- Large / complex state-space, action space.
- Long horizon
- Limited adaptivity (cannot run too many iterations)

Or alternatively, when offline data are provided, we can consider learning in the offline mode!

Offline Reinforcement Learning: doing policy optimization using historical data



Key question we ask: how to design efficient algorithm to reduce sample complexity?

Overview of the results

1. Propose offline RL algorithm for tabular MDPs [YW21]:
   - Under partial coverage assumption
   - Nearly-tight complexity:

$$\tilde{O}\left(\sum_{h=1}^{H}\sum_{s_h,a_h} d_h^{\pi^*}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^{*}+r_h)}{d_h^{\mu}(s_h,a_h)}}\cdot\sqrt{\frac{1}{n}}\right)$$

2. Propose offline RL algorithm for linear MDPs [YDWW22]:
   - Under the minimal eigenvalue condition
   - Instance-dependent guarantee (via variance-aware pessimistic learning)

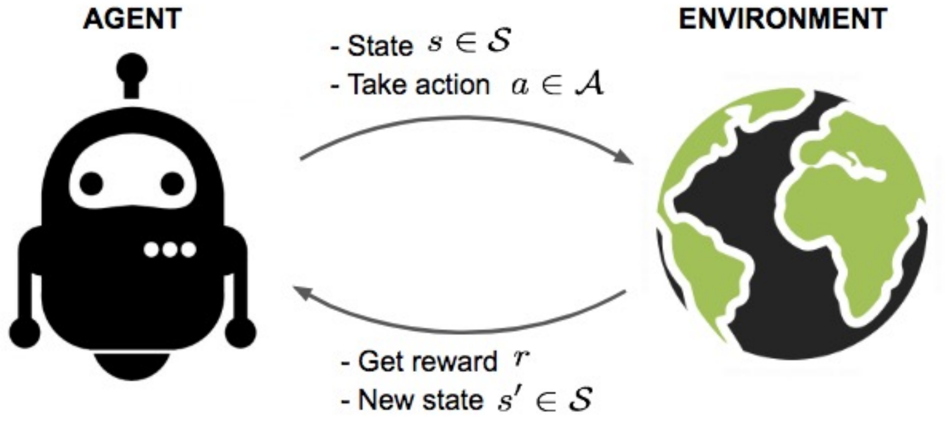
Outline

1. Overview of our works
2. Preliminary
3. Tabular Setting
4. Linear MDP setting
5. Summary

Offline learning in finite-horizon time-inhomogeneous MDPs

- Offline setting: batch data $\mathcal{D} = \{(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)}, r_t^{(i)})\}$, $t = 1, \ldots, H; i = 1, \ldots, n.$

Assuming behavior policy μ



Objective:

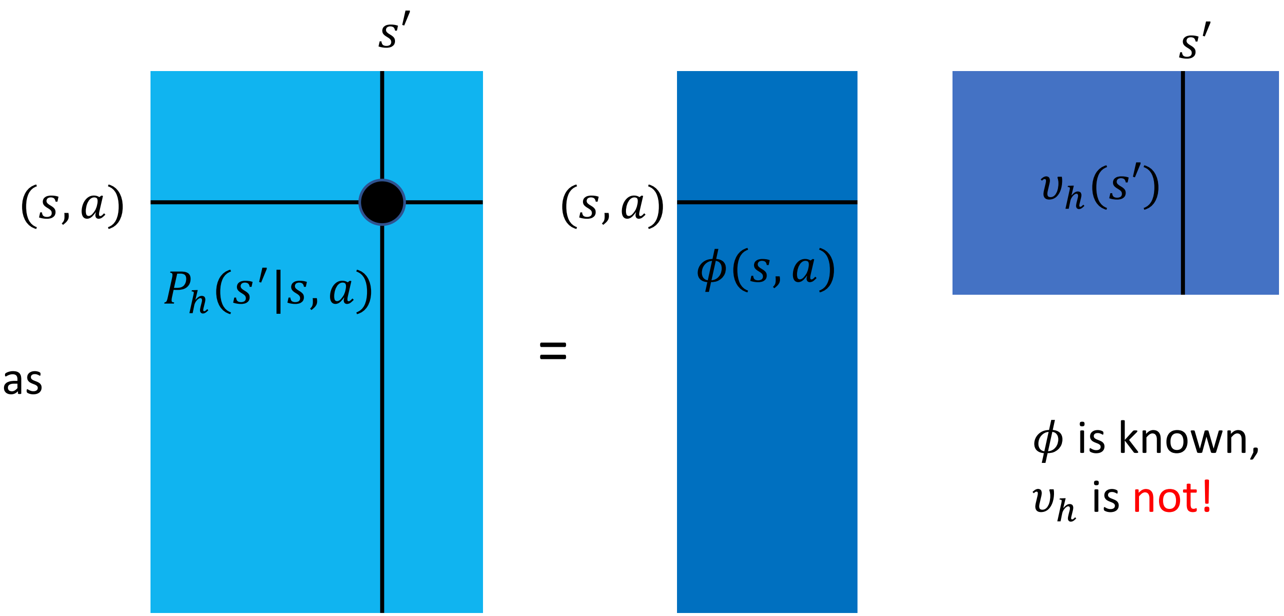
$$\max_{\pi} v^{\pi} := \mathbb{E}[\Sigma_{t=1}^{H} r(s_t, a_t) | a_t \sim \pi_t, P_1, \ldots, P_H]$$

Tabular setting

Discrete MDPs with finite states and actions

Linear MDP setting

Transition matrix P_h has rank d



ϕ is known,
v_h is not!

$$\exists \mu, \phi: \quad \forall s, a, s', P_h(s'|s,a) = \nu^T(s')\phi(s,a), \nu(\cdot) \in \mathbb{R}^d, \phi(\cdot,\cdot) \in \mathbb{R}^d$$

- Linear MDPs [YW20; JYWJ20] has low-rank structure, can generalize over infinite state action spaces;
- Relate to other models: e.g. low-rank MDPs [AKKS20; UZS22]
- Extensively studied in online setting, e.g. [DQC21; DJQ21]

Previous sample complexity results in offline learning

	Sample Complexity	Assumption	Setting
DVR[YBW21]	$\tilde{O}(H^3/d_m\varepsilon^2)$	Uniform coverage	Tabular
PEVI-ADV [XHWXB21;RZMJR21]	$\tilde{O}(H^3SC^*/\varepsilon^2)$	Single Concentrability	Tabular
Model-free[SLWCC22]	$\tilde{O}(H^3SC^*/\varepsilon^2)$	Single Concentrability	Tabular
PVI[JYW21]	$dH\Sigma_{h=1}^{H}\mathbb{E}_{\pi^*}[\|\|\phi(s_h,a_h)\|\|_{\Lambda_h^{-1}}]$	Compliance	Linear MDP
Bellman-Pessimism[XCJMA21]	$\sqrt{\frac{(1-\gamma)^{-4}d}{n}}\mathbb{E}_{\pi}[\|\|\phi(s,a)\|\|_{\Sigma_D^{-1}}]$	Realizability+ Completeness	Linear MDP
PACLE[ZWB21]	$\sqrt{d}\Sigma_{h=1}^{H}\ [\|\|\mathbb{E}_{\pi^*}\phi(s_h,a_h)\|\|_{\Sigma_h^{-1}}]$	Bellman Restricted Closedness	Linear MDP

Outline

1. Overview of our works
2. Preliminary
3. Tabular Setting [Yin&Wang21]
4. Linear MDP setting
5. Summary

We will *not* deal with exploration in offline RL, because we can't: assumption needed

- Uniform data coverage:
  - $d_m \coloneqq \min_{h,s_h,a_h} d_h^{\mu}(s_h, a_h) > 0$,
  - d_h^{μ} is the marginal state-action distribution.

- Uniform concentrability:
  - $C_\mu \coloneqq \sup_{\pi,h} \left\| \frac{d_h^{\pi}(\cdot,\cdot)}{d_h^{\mu}(\cdot,\cdot)} \right\| < \infty$.

- Single concentrability:
  - There exists π^* s.t. $d_h^{\mu}(s_h, a_h) > 0$ if $d_h^{\pi^*}(s_h, a_h) > 0$.

- What if no assumption is made about μ?

We will *not* deal with exploration in offline RL, because we can't: assumption needed

- Uniform data coverage (Assumption 2.1):
  - $d_m := \min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$,
  - d_h^μ is the marginal state-action distribution.

$$\varepsilon \approx \sqrt{\frac{H^3}{n \cdot d_m}}$$

[Yin, Bai, Wang, 2021]

- Uniform concentrability (Assumption 2.2):
  - $C_\mu := \sup_{\pi,h} \left\| \frac{d_h^\pi(\cdot,\cdot)}{d_h^\mu(\cdot,\cdot)} \right\| < \infty$.

- Single concentrability (Assumption 2.3):
  - There exists π^* s.t. $d_h^\mu(s_h, a_h) > 0$ if $d_h^{\pi^*}(s_h, a_h) > 0$.
  - The single concentrability $C^* = \max_{s,a} \frac{d_h^{\pi^*}(s,a)}{d_h^\mu(s,a)}$.

$$\varepsilon \approx \sqrt{\frac{H^3 S C^*}{n}}$$

[Xie et al., 2021]

- What if no assumption is made about μ?
  - Might suffer constant suboptimality gap.

Our Algorithm

Recap: UCBVI in Online RL

UCBVI [Azar et al. 2017]

- For $k = 0, \ldots, K-1$
- For $h = 1, \ldots, H$
  - Compute empirical estimate $\widehat{P}_h^k$;
  - Value Iteration with Optimism:
  - $\widehat{Q}_h^k(s,a) = \min\{r_h + \widehat{P}_h^k \widehat{V}_{h+1}^k + \Gamma_h^k, H - h + 1\}$,
  - $\widehat{V}_h^k(s) = \max_a \widehat{Q}_h^k(s,a)$,
  - $\hat{\pi}_h(s) = \mathrm{argmax}_a \widehat{Q}_h^k(s,a)$.

UCBVI vs. LCBVI, Online RL vs. Offline RL

UCBVI [Azar et al. 2017]

- For $k = 0, \dots, K-1$
- For $h = 1, \dots, H$
  - Compute empirical estimate $\hat{P}_h^k$;
  - Value Iteration with Optimism:
  - $\hat{Q}_h^k(s,a) = \min\{r_h + \hat{P}_h^k \hat{V}_{h+1}^k + \Gamma_h^k, H - h + 1\}$,
  - $\hat{V}_h^k(s) = \max_a \hat{Q}_h^k(s,a)$,
  - $\hat{\pi}_h(s) = \mathrm{argmax}_a \hat{Q}_h^k(s,a)$.

LCBVI ([Yin&Wang,21])

- For h= $H, \dots, 1$, use batch data
  - Compute empirical estimate $\hat{P}_h$;
  - Value Iteration with Pessimism:
  - $\hat{Q}_h(s,a) = \min\{r_h + \hat{P}_h \hat{V}_{h+1} - \Gamma_h, H - h + 1\}$,
  - $\hat{V}_h(s) = \max_a \hat{Q}_h(s,a)$,
  - $\hat{\pi}_h(s) = \mathrm{argmax}_a \hat{Q}_h(s,a)$.

The design of bonus Γ_h matters!

LCBVI-Bernstein: Adaptive Pessimistic Value Iteration, simple algorithm ☺

For h= $H, \dots, 1$, use batch data

- Compute empirical estimate $\widehat{P}_h$;
- Value Iteration with Pessimism:
- $\widehat{Q}_h(s,a) = \min\{r_h + \widehat{P}_h \widehat{V}_{h+1} - \Gamma_h, H - h + 1\}_+$,
- $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$,
- $\hat{\pi}_h(s) = \mathrm{argmax}_a \widehat{Q}_h(s,a)$.

Insert $\Gamma_h(s_h, a_h) \approx \sqrt{\frac{\mathrm{Var}_{\widehat{P}_{s_h,a_h}}(\hat{r}_h + \widehat{V}_{h+1})}{n_{s_h,a_h}}} + \frac{H}{n_{s_h,a_h}}$ if $n_{s_h,a_h} \geq 1$, o.w. $\frac{CH}{1}$.

As a result: APVI/LCBVI-Bernstein gives intrinsic offline reinforcement learning bound

$$0 \leq v^* - v^{\widehat{\pi}} \leq C \sum_{h=1}^{H} \sum_{s_h, a_h} d_h^{\pi^*}(s_h, a_h) \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^* + r_h)}{d_h^{\mu}(s_h, a_h)}} \cdot \sqrt{\frac{1}{n}} + O(\frac{H^3}{n d_m})$$

- Directly implication of the intrinsic offline RL bound:
  - Under Uniform data coverage: reduces to $O(\sqrt{\frac{H^3}{n d_m}})$, near-minimax optimal [**Yin** et al. 2021a];

As a result: APVI/LCBVI-Bernstein gives intrinsic offline reinforcement learning bound

$$0 \le v^* - v^{\widehat{\pi}} \le C \sum_{h=1}^{H} \sum_{s_h, a_h} d_h^{\pi^*}(s_h, a_h) \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^* + r_h)}{d_h^{\mu}(s_h, a_h)}} \cdot \sqrt{\frac{1}{n}} + O(\frac{H^3}{n d_m})$$

- Directly implication of the intrinsic offline RL bound:
  - Under Uniform data coverage: reduces to $O(\sqrt{\frac{H^3}{n d_m}})$, near-minimax optimal [**<u>Yin</u>** et al. 2021a];
  - Single concentrability: reduces to $O\left(\sqrt{\frac{H^3 S C^*}{n}}\right)$, near-minimax optimal [Xie et al. 2021b];

As a result: APVI/LCBVI-Bernstein gives intrinsic offline reinforcement learning bound

$$0 \leq v^* - v^{\widehat{\pi}} \leq C \sum_{h=1}^{H} \sum_{s_h, a_h} d_h^{\pi^*}(s_h, a_h) \sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^* + r_h)}{d_h^{\mu}(s_h, a_h)}} \cdot \sqrt{\frac{1}{n}} + O(\frac{H^3}{n d_m})$$

- Directly implication of the intrinsic offline RL bound:
  - Under Uniform data coverage: reduces to $O(\sqrt{\frac{H^3}{n d_m}})$, near-minimax optimal [**Yin** et al. 2021a];
  - Single concentrability: reduces to $O\left(\sqrt{\frac{H^3 S C^*}{n}}\right)$, near-minimax optimal [Xie et al. 2021b];
  - Problem-dependent expression: $\tilde{O}\left(\sum_{h=1}^{H} \sqrt{\frac{\mathbb{Q}_h^*}{n \cdot d_m}}\right) + \tilde{O}\left(\frac{H^3}{n \cdot d_m}\right)$, mirrors [Zanette and Brunskill, 2019].

A bit more on problem-dependent domain

- Intrinsic bound can be simplified to the following by denoting $\mathbb{Q}_h^* := \min_{s,a} Var_{P_{s,a}}(V_{h+1}^* + r_h)$:

$$\tilde{O}\left(\sum_{h=1}^{H}\sqrt{\frac{\mathbb{Q}_h^*}{n \cdot d_m}}\right) + \tilde{O}\left(\frac{H^3}{n \cdot d_m}\right)$$

  - Deterministic systems: when $\mathbb{Q}_h^* = 0$, it automatically yields faster convergence

$$\Rightarrow \quad \hat{O}\left(\frac{H^3}{n \cdot d_m}\right)$$

  - Partially deterministic (mixture) systems: t stage stochastic transitions and $H - t$ stage deterministic transitions

$$\Rightarrow \quad t \cdot \sqrt{\max_h \mathbb{Q}_h^* / n\bar{d}_m}$$

Everything in one figure…

Intrinsic Offline Learning Bound

$$\sum_{h=1}^{H}\sum_{s_h,a_h} d_h^{\pi^*}(s_h,a_h)\sqrt{\frac{\mathrm{Var}_{P_{s_h,a_h}}(V_{h+1}^{*}+r_h)}{d_h^{\mu}(s_h,a_h)}}\cdot\sqrt{\frac{1}{n}}$$

Uniform Visitation

$$\widetilde{O}\left(\sqrt{\frac{H^3}{n\cdot d_m}}\right)$$

Single Concentrability

$$\widetilde{O}\left(\sqrt{\frac{H^3 SC^*}{n}}\right)$$

Adaptive Domain

$$\widetilde{O}\left(\sum_{h=1}^{H}\sqrt{\frac{\mathbb{Q}_h^*}{n\cdot d_m}}\right)+\widetilde{O}\left(\frac{H^3}{n\cdot d_m}\right)$$

How to certify this is near-optimal (at instance level)?

- We also have
  - An instance-dependent lower bound (Theorem 4.3);
  - Assumption-Free RL (Theorem 5.1)
  - …

In particular, we need to leverage the variance structure to create local hard instance for every transition

$$P_h'(s'|s,a) = P_h(s'|s,a) + \frac{P_h(s'|s,a)(V_{h+1}^*(s_{h+1}) - \mathrm{E}_P[V_{h+1}^*])}{\sqrt{\xi \cdot n_{s,a} \cdot Var_{P_{s,a}}(V_{h+1}^*)}}$$

What give rise to instance-dependent structure?

Leveraging Extended Value Difference Lemma

$$v^* - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(s_h, a_h)] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\xi_h(s_h, a_h)]$$

What give rise to instance-dependent structure?

Leveraging Extended Value Difference Lemma

$$v^* - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(s_h, a_h)] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\xi_h(s_h, a_h)]$$

Leveraging Empirical Bernstein inequality

$$\xi_h(s_h, a_h) \lesssim \sqrt{\frac{Var_{\hat{P}}(\hat{r} + \widehat{V}_{h+1})}{n_{s_h, a_h}}}$$

What give rise to instance-dependent structure?

Leveraging Extended Value Difference Lemma

$$v^* - v^{\widehat{\pi}} \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\xi_h(s_h, a_h)] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\xi_h(s_h, a_h)]$$

Leveraging Empirical Bernstein inequality

$$\xi_h(s_h, a_h) \lesssim \sqrt{\frac{Var_{\hat{P}}(\hat{r} + \widehat{V}_{h+1})}{n_{s_h, a_h}}}$$

Converting sample-level quantities to population quantities

Outline

1. Overview of our works
2. Preliminary
3. Tabular Setting
4. Linear MDP setting [YDWW22]
5. Summary

Going beyond Tabular setting

- Well… there are works study linear MDPs
  - Pessimistic Value Iteration [JYW21]
  - Bellman-consistent Pessimism [XCJMA21]
  - Pessimistic Actor-Critic [ZWB21]
  - …
- But they are not tight in general (no matching bounds)

Is tighter instance-dependent bounds possible?

From the technical end

- Improvement could be challenging, since all previous analysis rely on the self-nomalized Hoeffding’s bound technique
- Has been exploited extensively since the online analysis [JYWJ20]

From the technical end

- Improvement could be challenging, since all previous analysis rely on the self-nomalized Hoeffding's bound technique
- Has been exploited extensively since the online analysis [JYWJ20]

Good news

- [ZGS21] introduced the self-nomalized Bernstein's bound technique to obtain the near-optimal regret for linear mixture MDPs
- Has been successfully applied to the linear MDP OPE problem [MWZG21]

Also, what is missing?

- Previous algorithms consider least-square value iteration objective

$$\widehat{w}_h := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|w\|_2^2 + \sum_{k=1}^{K} \left[\langle \phi(s_h^k, a_h^k), w \rangle - r_h^k - V_{h+1}(s_{h+1}'^k) \right]^2$$

- The “default” choice for linear-regression-type problems

Also, what is missing?

- Previous algorithms consider least-square value iteration objective

$$\widehat{w}_h := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|w\|_2^2 + \sum_{k=1}^{K} \Big[\langle \phi(s_h^k, a_h^k), w \rangle - r_h^k - V_{h+1}(s'^k_{h+1}) \Big]^2$$

- The “default” choice for linear-regression-type problems

- However, RL is more than that…

  - RL is heterogeneous in nature as different (s, a) corresponds to different distributions $P(\cdot \mid s, a)$

  - Intuitively, causes samples with low variance in transitions more informative than others

Also, what is missing?

Modification: better to reweight the LSVI objective according to their (estimated) uncertainties

Also, what is missing?

Modification: better to reweight the LSVI objective according to their (estimated) uncertainties

$$\widehat{w}_h := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|w\|_2^2 + \sum_{k=1}^{K} \left[\langle \phi(s_h^k, a_h^k), w \rangle - r_h^k - V_{h+1}(s_{h+1}'^k) \right]^2$$

LSVI

$$\widehat{w}_h := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \lambda \|w\|_2^2 + \sum_{k=1}^{K} \frac{\left[\langle \phi(s_h^k, a_h^k), w \rangle - r_h^k - \widehat{V}_{h+1}(s_{h+1}'^k) \right]^2}{\widehat{\sigma}_h^2(s_h^k, a_h^k)}$$

Weighted LSVI

Variance-Aware Pessimistic Value Iteration [YDWW22]

Algorithm 1 **V**ariance-**A**ware **P**essimistic **V**alue **I**teration (VAPVI)

Input: Dataset $\mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$ $\mathcal{D}' = \{(\bar{s}_h^\tau, \bar{a}_h^\tau, \bar{r}_h^\tau)\}_{\tau,h=1}^{K,H}$. Universal constant C.

Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.

for $h = H, H-1, \ldots, 1$ **do**

  Set $\bar{\Sigma}_h \leftarrow \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau)\phi(\bar{s}_h^\tau, \bar{a}_h^\tau)^\top + \lambda I$

  Set $\bar{\beta}_h \leftarrow \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau) \cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)^2$

  Set $\bar{\theta}_h \leftarrow \bar{\Sigma}_h^{-1} \sum_{\tau=1}^K \phi(\bar{s}_h^\tau, \bar{a}_h^\tau) \cdot \widehat{V}_{h+1}(\bar{s}_{h+1}^\tau)$

  Set $\left[\widehat{\mathrm{Var}}_h \widehat{V}_{h+1}\right](\cdot,\cdot) = \left\langle \phi(\cdot,\cdot), \bar{\boldsymbol{\beta}}_h \right\rangle_{[0,(H-h+1)^2]} - \left[\left\langle \phi(\cdot,\cdot), \bar{\boldsymbol{\theta}}_h \right\rangle_{[0,H-h+1]}\right]^2$

  Set $\widehat{\sigma}_h(\cdot,\cdot)^2 \leftarrow \max\{1, \widehat{\mathrm{Var}}_{P_h} \widehat{V}_{h+1}(\cdot,\cdot)\}$

  Set $\widehat{\Lambda}_h \leftarrow \sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right) \phi\left(s_h^\tau, a_h^\tau\right)^\top / \widehat{\sigma}^2(s_h^\tau, a_h^\tau) + \lambda \cdot I$,

  Set $\widehat{w}_h \leftarrow \widehat{\Lambda}_h^{-1} \left(\sum_{\tau=1}^K \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(r_h^\tau + \widehat{V}_{h+1}\left(s_{h+1}^\tau\right)\right) / \widehat{\sigma}^2(s_h^\tau, a_h^\tau)\right)$

  Set $\Gamma_h(\cdot,\cdot) \leftarrow C\sqrt{d} \cdot \left(\phi(\cdot,\cdot)^\top \widehat{\Lambda}_h^{-1} \phi(\cdot,\cdot)\right)^{1/2} + \frac{2H^3\sqrt{d}}{K}$ (Use Γ_h^I for the improved version)

  Set $\bar{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{w}_h - \Gamma_h(\cdot,\cdot)$

  Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\left\{\bar{Q}_h(\cdot,\cdot), H-h+1\right\}^+$

  Set $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot) \right\rangle_{\mathcal{A}}$, $\widehat{V}_h(\cdot) \leftarrow \max_{\pi_h} \left\langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot) \right\rangle_{\mathcal{A}}$

end for

Output: $\{\widehat{\pi}_h\}_{h=1}^H$.

Variance-Aware Pessimistic Value Iteration (VAPVI)

Our result for linear MDP

Under minimal eigenvalue condition $min_h \lambda_{min}(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^T] \coloneqq \kappa > 0)$

$$v^\star - v^{\widehat{\pi}} \leq \widetilde{O}\big(\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top \Lambda_h^{\star-1} \phi(\cdot,\cdot)}\right]\big) + \frac{2H^4\sqrt{d}}{K}$$

where $\Lambda_h^\star = \sum_{k=1}^{K} \frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^\top}{\sigma^2_{V^\star_{h+1}(s_h^k,a_h^k)}} + \lambda I_d$ *and* $\widetilde{O}$ *hides universal constants and the Polylog terms.*

Variance-Aware Pessimistic Value Iteration (VAPVI)

Our result for linear MDP

Under minimal eigenvalue condition $min_h \lambda_{min}(\mathbb{E}_{\mu,h}[\phi(s,a)\phi(s,a)^T] := \kappa > 0)$

$$v^\star - v^{\widehat{\pi}} \leq \widetilde{O}\big(\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\sqrt{\phi(\cdot,\cdot)^\top \Lambda_h^{\star -1} \phi(\cdot,\cdot)}\right]\big) + \frac{2H^4\sqrt{d}}{K}$$

where $\Lambda_h^\star = \sum_{k=1}^{K} \frac{\phi(s_h^k,a_h^k)\cdot\phi(s_h^k,a_h^k)^\top}{\sigma^2_{V_{h+1}^\star(s_h^k,a_h^k)}} + \lambda I_d$ *and* $\widetilde{O}$ *hides universal constants and the Polylog terms.*

In addition, the output policy $\hat{\pi}$ can compete with any policy!

Comparison with previous results

Pessimistic value iteration
[Jin et al. 2021]

$$dH\Sigma_{h=1}^{H}\mathbb{E}_{\pi^*}[||\phi(s_h,a_h)||_{\Sigma_h^{-1}}]$$

$\approx \sqrt{d}$

$\Sigma_h \approx \Sigma_{k=1}^{K}\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^T + \lambda I$

Bellman-consistent
(linear MDP result)
[Xie et al. 2021]

$$\sqrt{\frac{(1-\gamma)^{-4}d}{n}}\mathbb{E}_{\pi^*}[||\phi(s,a)||_{\Sigma_D^{-1}}]$$

H^2 to $\sigma^2_{V^*_{h+1}}$

Variance-aware
pessimism (ours)

$\Lambda_h \approx \Sigma_{k=1}^{K}\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^T/\sigma^2_{V^*_{h+1}} + \lambda I$

$$\sqrt{d}\Sigma_{h=1}^{H}\mathbb{E}_{\pi^*}[||\phi(s_h,a_h)||_{\Lambda_h^{-1}}]$$

What's more

- Preserves instance-dependent features

  e.g. when the instance has deterministic system, ensures faster convergence $\frac{2H^4\sqrt{d}}{K}$

- The guarantee can be further improved if non-negative feature is given ($\phi \geq 0$)

$$\widetilde{O}\left(\sqrt{d} \cdot \sum_{h=1}^{H} \sqrt{\mathbb{E}_\pi[\phi(\cdot,\cdot)]^\top \Lambda_h^{-1} \mathbb{E}_\pi[\phi(\cdot,\cdot)]}\right).$$

- Improvement is strict when reduce to tabular setting!

- Self-normalized Bernstein inequality is the key for improvement!

Summary

- For both tabular and linear MDP setting, we provide get tighter instance-dependent bounds
  - For the tabular case, it subsumes previous worst-case bounds
  - For the linear MDP case, can incorporate variance structure and improve over the previous results
- Future Directions
  - Weaken the minimal eigenvalue assumption for linear MDPs
  - Extending to more general function approximation setting (e.g. differentiable function classes)

Summary

- Based on

  - Towards Instance-Optimal Offline Reinforcement Learning with Pessimism [Yin Ming&Wang Yu-Xiang, NeurIPS21]

  - Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism [Yin Ming, Duan Yaqi, Wang Mengdi, Wang Yu-Xiang, ICLR22]

Thank you!