Is Pessimism Provably Efficient for Offline RL?

Ying Jin
Stanford Statistics

RL Theory Virtual Seminars
April 6, 2020

Joint work with

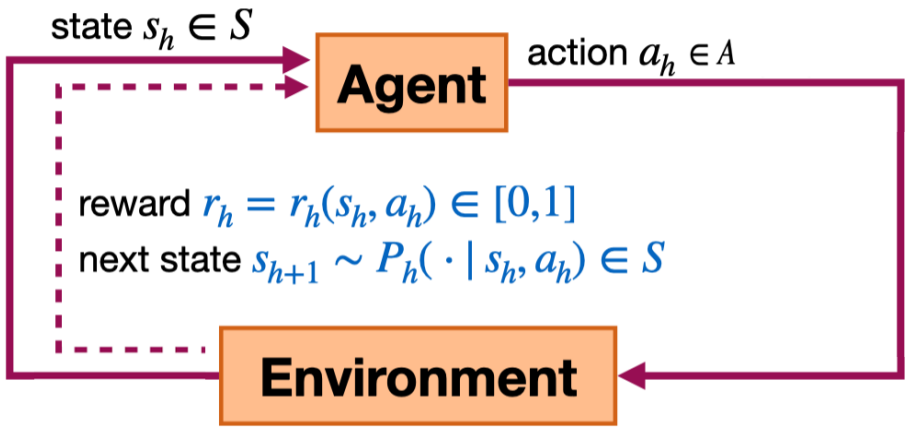
Zhuoran Yang
Princeton ORFE

Zhaoran Wang
Northwestern IEMS

Backgrounds

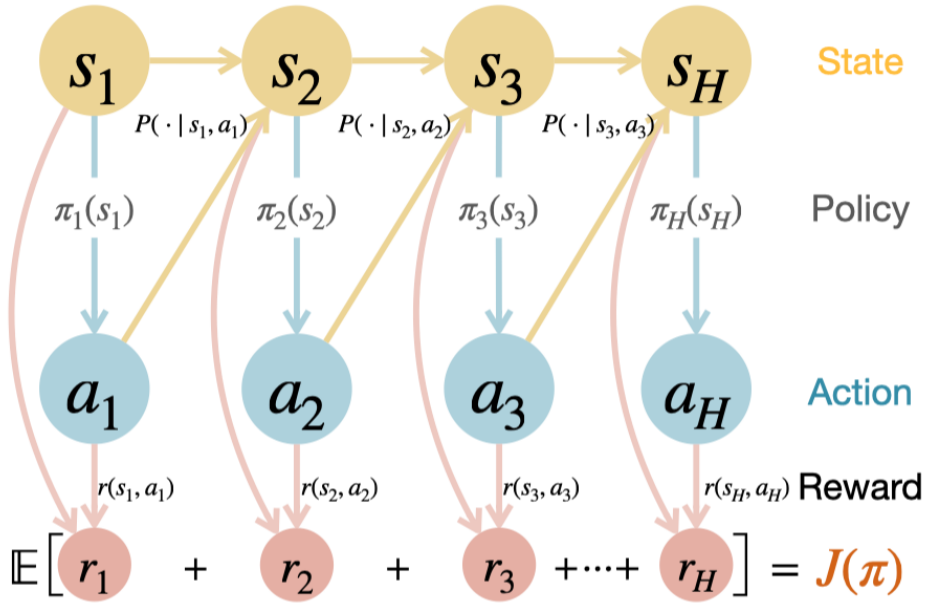
- Episodic MDP
- Offline Learning Protocol
- Difficulty and Epistemic Uncertainty

Episodic MDP



- $\mathcal{S}$: infinite state space. $\mathcal{A}$: finite action space.
- Unknown reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$.
- Unknown transition kernel $\mathbb{P}_h(\cdot \mid x, a) \in \Delta(\mathcal{S})$.
- Finite horizon H: terminate when $h = H$.

Episodic MDP



- Policy: $\pi = \{\pi_h\}_{h \in [H]} : \mathcal{S} \to \Delta(\mathcal{A})$, $a_h \sim \pi_h(s_h)$.
- Expected total reward: $J(\pi, x) = \mathbb{E}_\pi[\sum_{h=1}^{H} r_h \mid s_1 = x] \in [0, H]$.
- Optimal policy: $\pi^\star(\cdot) = \mathrm{argmax}_\pi \, J(\pi, \cdot)$.

Episodic MDP

Dynamic Programming and Bellman Equation

- $\pi^\star$ is greedy w.r.t. optimal value functions $Q^\star = \{Q_h^\star\}_{h\in[H]}$.

Episodic MDP

Dynamic Programming and Bellman Equation

- $\pi^\star$ is greedy w.r.t. optimal value functions $Q^\star = \{Q_h^\star\}_{h\in[H]}$.

$$\pi_h^\star(x) = \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^\star(x,a), \quad \forall s \in S,$$

$$Q_h^\star(x,a) = \underbrace{r_h(x,a) + \mathbb{E}_{s_{h+1}\sim\mathbb{P}}\Big[\max_{a'\in\mathcal{A}} Q_{h+1}^\star(s_{h+1},a') \,\Big|\, s_h = x, a_h = a\Big]}_{\text{Bellman operator } \mathbb{B}_h\colon\ \mathbb{B}_h Q_{h+1}^\star}$$

Episodic MDP

Dynamic Programming and Bellman Equation

- $\pi^\star$ is greedy w.r.t. optimal value functions $Q^\star = \{Q_h^\star\}_{h\in[H]}$.

$$\pi_h^\star(x) = \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^\star(x,a), \quad \forall s \in S,$$

$$Q_h^\star(x,a) = \underbrace{r_h(x,a) + \mathbb{E}_{s_{h+1}\sim\mathbb{P}}\Big[\max_{a'\in\mathcal{A}} Q_{h+1}^\star(s_{h+1},a') \,\Big|\, s_h = x, a_h = a\Big]}_{\text{Bellman operator } \mathbb{B}_h\colon\ \mathbb{B}_h Q_{h+1}^\star}$$

- Bellman Equation

$$Q_h^\star = \mathbb{B}_h Q_{h+1}^\star, \quad Q_{H+1}^\star \equiv 0.$$

Episodic MDP

Dynamic Programming and Bellman Equation

- $\pi^\star$ is greedy w.r.t. optimal value functions $Q^\star = \{Q_h^\star\}_{h\in[H]}$.

$$\pi_h^\star(x) = \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^\star(x,a), \quad \forall s \in S,$$

$$Q_h^\star(x,a) = \underbrace{r_h(x,a) + \mathbb{E}_{s_{h+1}\sim\mathbb{P}}\Big[\max_{a'\in\mathcal{A}} Q_{h+1}^\star(s_{h+1},a') \,\Big|\, s_h = x, a_h = a\Big]}_{\text{Bellman operator } \mathbb{B}_h\colon\ \mathbb{B}_h Q_{h+1}^\star}$$

- Bellman Equation

$$Q_h^\star = \mathbb{B}_h Q_{h+1}^\star, \quad Q_{H+1}^\star \equiv 0.$$

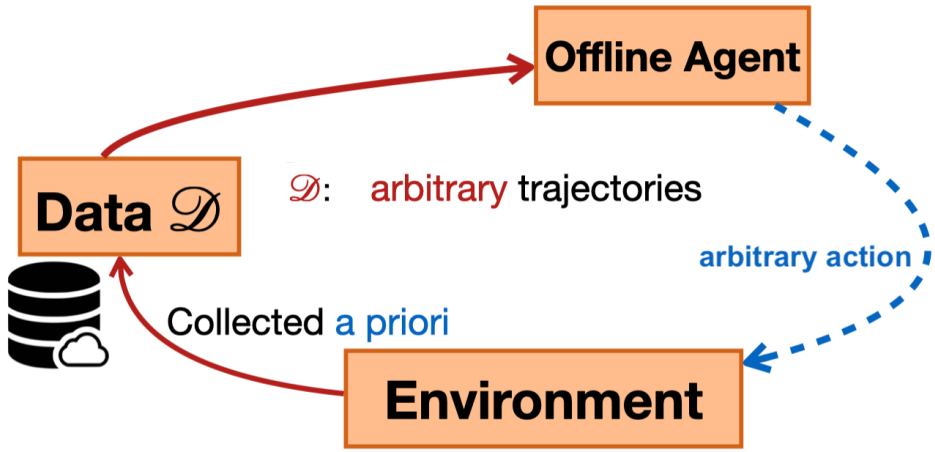
- RL with function approximation:

Function class $\mathcal{F} = \{f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}\}$ approximates $Q_h^\star$.

Linear functions, Neural networks, RKHS...

Offline Policy Learning

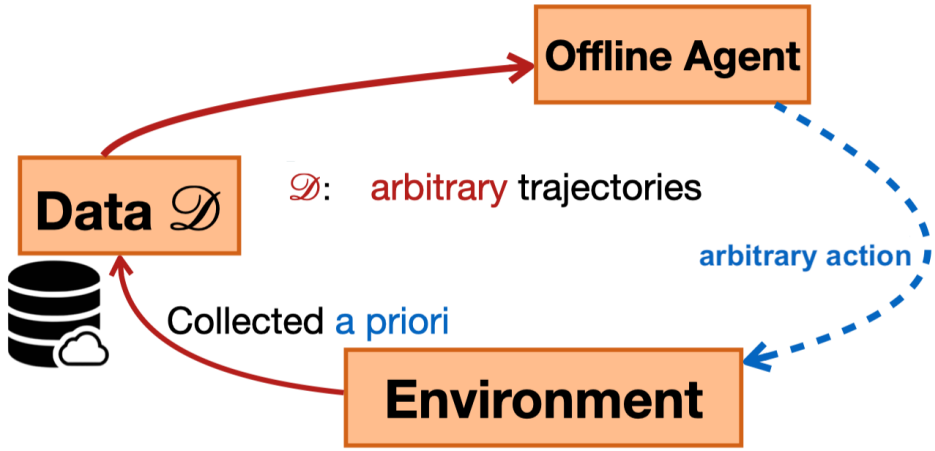
Learn from Given Datasets



- Offline Data: collected a priori.
- Arbitrary trajectories: actions a_h by an offline agent (unknown rule).
- No further interactions with MDP.

Offline Policy Learning

Learn from Given Datasets



- Offline Data: collected a priori.
- Arbitrary trajectories: actions a_h by an offline agent (unknown rule).
- No further interactions with MDP.
- Learning objective: performance of the learned policy

$$\text{SubOpt}(\widehat{\pi}, x) = J(\pi^\star, x) - J(\widehat{\pi}, x),$$

where $\widehat{\pi}$ =OfflineRL$(\mathcal{D}, \mathcal{F})$, $x \in \mathcal{S}$.

Offline Policy Learning

Naive Value Iterations

- By Bellman equation: approximate dynamic programming.

Naive Value Iterations

- End of Episode: $\widehat{Q}_{H+1} \leftarrow 0$.
- Dynamic Programming: $h = H, H-1, \ldots, 1$,
  - Estimate: $\widehat{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
  - Optimize: $\widehat{\pi}_h(x) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x, a)$.

Offline Policy Learning

Naive Value Iterations

- By Bellman equation: approximate dynamic programming.

Naive Value Iterations

- End of Episode: $\widehat{Q}_{H+1} \leftarrow 0$.
- Dynamic Programming: $h = H, H-1, \ldots, 1$,
  - Estimate: $\widehat{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
  - Optimize: $\widehat{\pi}_h(x) \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x, a)$.

- Hope for ...
  - Good estimation: $\widehat{Q}_h \approx Q_h^\star$.
  - Good optimization:

$$\operatorname*{argmax}_a \widehat{Q}_h(x, a) \approx \operatorname*{argmax}_a Q_h^\star(x, a).$$

  - Good actual performance:

$$J(\widehat{\pi}) \approx J(\pi^\star).$$

Why May Naive Value Iterations Fail?

An Example: Multi-Arm Bandit

- Gaussian bandits

$$R_i \,|_{A_i=a} \sim N(r(a), 1), \quad a \in [K] \text{ arms.}$$

- a is pulled $N(a)$ times in the pre-collected dataset $\mathcal{D} = \{(A_i, R_i)\}$.

Why May Naive Value Iterations Fail?

An Example: Multi-Arm Bandit

- Gaussian bandits

$$R_i \,|_{A_i=a} \sim N(r(a), 1), \quad a \in [K] \text{ arms.}$$

- a is pulled $N(a)$ times in the pre-collected dataset $\mathcal{D} = \{(A_i, R_i)\}$.
- Following the naive value iterations,
  - Estimate by sample mean: $\widehat{\mu}(a) = \frac{\sum_{i:A_i=a} R_i}{N(a)}$.
  - Optimize by greedy: $\widehat{a} = \text{argmax}_{a\in[K]}\ \widehat{\mu}(a)$.

Why May Naive Value Iterations Fail?

An Example: Multi-Arm Bandit

- Gaussian bandits

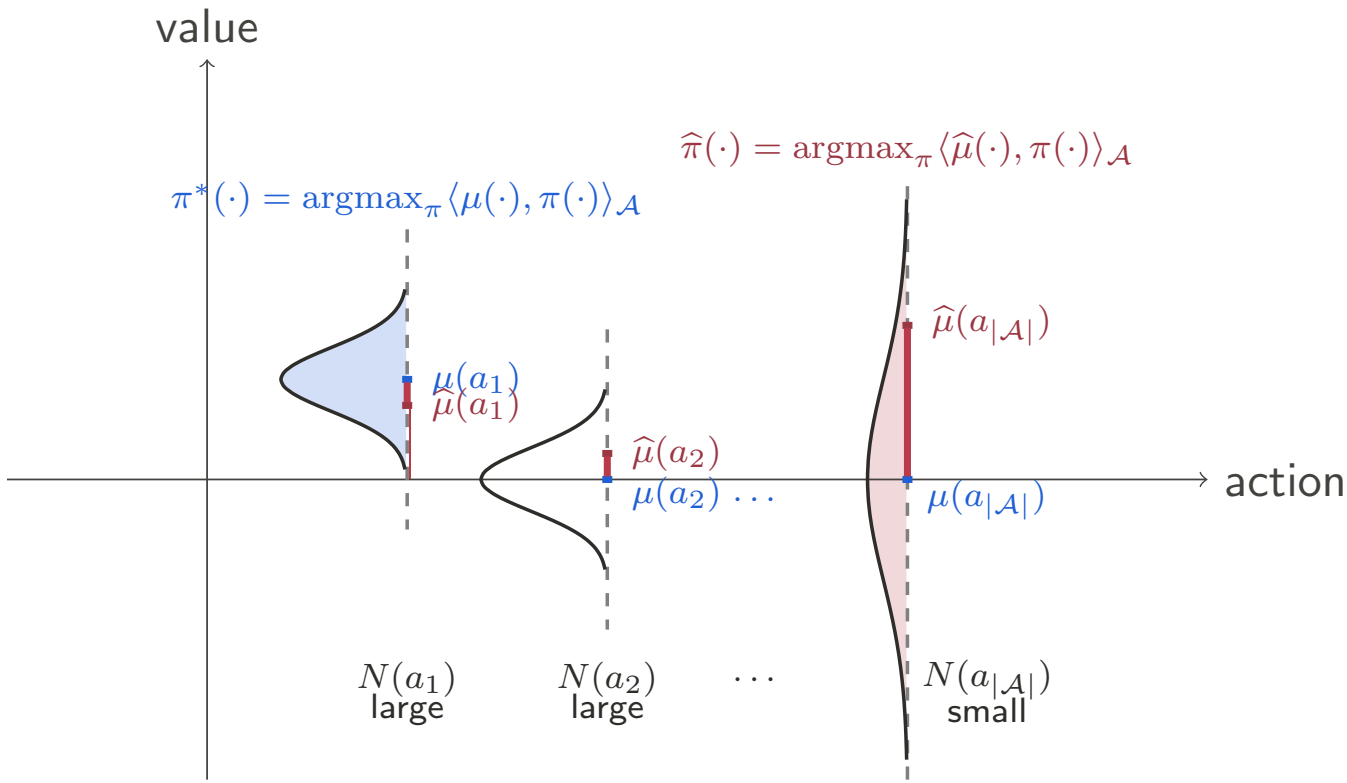
$$R_i \,|_{\,A_i=a} \sim N(r(a), 1), \quad a \in [K] \text{ arms.}$$

- a is pulled $N(a)$ times in the pre-collected dataset $\mathcal{D} = \{(A_i, R_i)\}$.
- Following the naive value iterations,
  - Estimate by sample mean: $\widehat{\mu}(a) = \frac{\sum_{i:A_i=a} R_i}{N(a)}$.
  - Optimize by greedy: $\widehat{a} = \operatorname{argmax}_{a\in[K]} \widehat{\mu}(a)$.
- For small $N(a)$, a bad arm a **might appear good by chance**.

Why May Naive Value Iterations Fail?

An Example: Multi-Arm Bandit

- For small $N(a)$, a bad arm a **might appear good by chance**.



Why May Naive Value Iterations Fail?

Epistemic Uncertainty

- Policy $\widetilde{\pi}$ insufficiently covered by dataset $\mathcal{D}$
  $\Rightarrow$ Large uncertainty in our knowledge about a policy $\widetilde{\pi}$.

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

- Policy $\widetilde{\pi}$ insufficiently covered by dataset $\mathcal{D}$

  $\Rightarrow$ Large uncertainty in our knowledge about a policy $\widetilde{\pi}$.

- Epistemic Uncertainty spuriously correlates with decision-making (greedy step). Roughly,

$$J(\widehat{\pi}) = J\big(\operatorname*{argmax}_{\pi} \widehat{J}(\pi)\big).$$

  $\widehat{J}$ might be far from J for some π.

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

- Policy $\widetilde{\pi}$ insufficiently covered by dataset $\mathcal{D}$

  $\Rightarrow$ Large uncertainty in our knowledge about a policy $\widetilde{\pi}$.

- Epistemic Uncertainty spuriously correlates with decision-making (greedy step). Roughly,

$$J(\widehat{\pi}) = J\big(\operatorname*{argmax}_{\pi} \widehat{J}(\pi)\big).$$

  $\widehat{J}$ might be far from J for some π.

- **Ruined if a bad π with large uncertainty appears to be good!**

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

- Policy $\widetilde{\pi}$ insufficiently covered by dataset $\mathcal{D}$

  $\Rightarrow$ Large uncertainty in our knowledge about a policy $\widetilde{\pi}$.

- Epistemic Uncertainty spuriously correlates with decision-making (greedy step). Roughly,

$$J(\widehat{\pi}) = J\big(\operatorname*{argmax}_{\pi} \widehat{J}(\pi)\big).$$

  $\widehat{J}$ might be far from J for some π.

- **Ruined if a bad π with large uncertainty appears to be good!**
- No further interactions with MDP $\Rightarrow$ unable to reduce uncertainty.

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

- **Ruined if a bad π with large uncertainty appears to be good!**
- No further interactions with MDP $\Rightarrow$ unable to reduce uncertainty.

- Uniform coverage for all policies? Too strong & unrealistic.

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

- Our solution by **Pessimism**: penalize large epistemic uncertainties.

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

- Our solution by **Pessimism**: penalize large epistemic uncertainties.
  - High estimated value, high uncertainty ☹

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

- Our solution by **Pessimism**: penalize large epistemic uncertainties.
  - High estimated value, high uncertainty ☹
  - High estimated value, low uncertainty 😄

Why May Naive Value Iterations Fail?

Epistemic Uncertainty

Question

Is it possible to design a provably efficient algorithm for offline RL under minimal assumptions on the dataset?

- Our solution by **Pessimism**: penalize large epistemic uncertainties.
  - High estimated value, high uncertainty ☹
  - High estimated value, low uncertainty 😄
  - Good estimated value, low uncertainty 😄

A Solution:
Pessimism for Offline Learning

- Pessimistic Value Iteration
- Why Pessimism Helps
- PEVI for Linear MDP

Pessimism for Offline Learning

Pessimism Principle

- Empirical success in practice:
  - Pessimistic model-based [Yu et al. (2020); Kidambi et al. (2020)]
  - Pessimistic value-based [Kumar et al. (2020)]
- In theory:
  - Regularized fitted Q-iterations [Liu et al. (2020)]: restrict policy class to be close to behavior policy.
  - Importance of Pessimism [Buckman et al. (2020)]

Pessimism for Offline Learning

Pessimism Principle

- Empirical success in practice:
  - Pessimistic model-based [Yu et al. (2020); Kidambi et al. (2020)]
  - Pessimistic value-based [Kumar et al. (2020)]
- In theory:
  - Regularized fitted Q-iterations [Liu et al. (2020)]: restrict policy class to be close to behavior policy.
  - Importance of Pessimism [Buckman et al. (2020)]

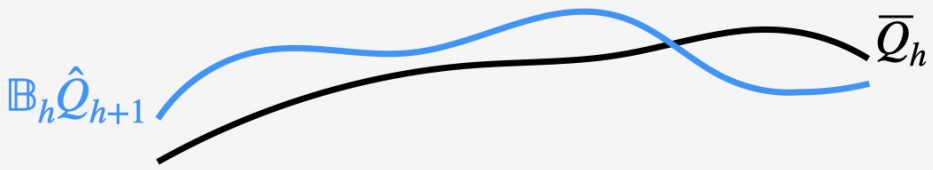
- This work: Pessimistic Value Iteration.
  - A principled framework for pessimism in value iterations.
  - No restriction on policy class and coverage of dataset.
  - Optimality of pessimism in the sense of information theory.

Pessimism for Offline Learning

General Algorithm: Pessimistic Value Iteration

Algorithm: Pessimistic Value Iterations (General Form)

- Estimate: $\overline{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.



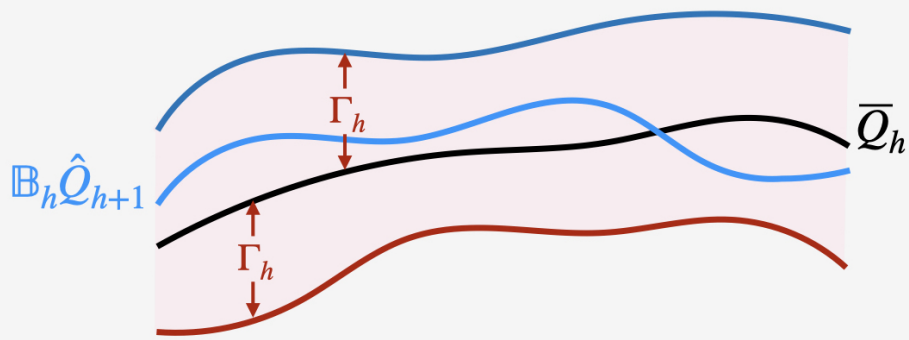
Pessimism for Offline Learning

General Algorithm: Pessimistic Value Iteration

Algorithm: Pessimistic Value Iterations (General Form)

- Estimate: $\overline{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
- Uncertainty quantification (UQ): w.h.p.

$$\left|\overline{Q}_h - (\mathbb{B}_h \widehat{Q}_{h+1})\right| \leq \Gamma_h, \quad \forall h \in [H].$$



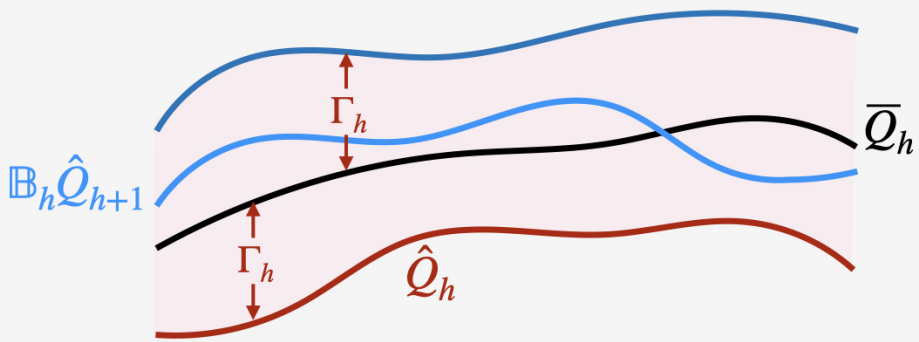
Pessimism for Offline Learning

General Algorithm: Pessimistic Value Iteration

Algorithm: Pessimistic Value Iterations (General Form)

- Estimate: $\overline{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
- Uncertainty quantification (UQ): w.h.p. $\left|\overline{Q}_h - (\mathbb{B}_h \widehat{Q}_{h+1})\right| \leq \Gamma_h$
- Construct pessimistic value function

$$\widehat{Q}_h(x,a) = \underbrace{\overline{Q}_h(x,a)}_{\text{VI}} \underbrace{-\Gamma_h(x,a)}_{\text{penalty}}$$



Pessimism for Offline Learning

General Algorithm: Pessimistic Value Iteration

Algorithm: Pessimistic Value Iterations (General Form)

- Estimate: $\overline{Q}_h \leftarrow \texttt{Regress}(\mathbb{B}_h \widehat{Q}_{h+1}, \mathcal{D}, \mathcal{F})$.
- Uncertainty quantification (UQ): w.h.p.

$$\left|\overline{Q}_h - (\mathbb{B}_h \widehat{Q}_{h+1})\right| \leq \Gamma_h, \quad \forall h \in [H].$$

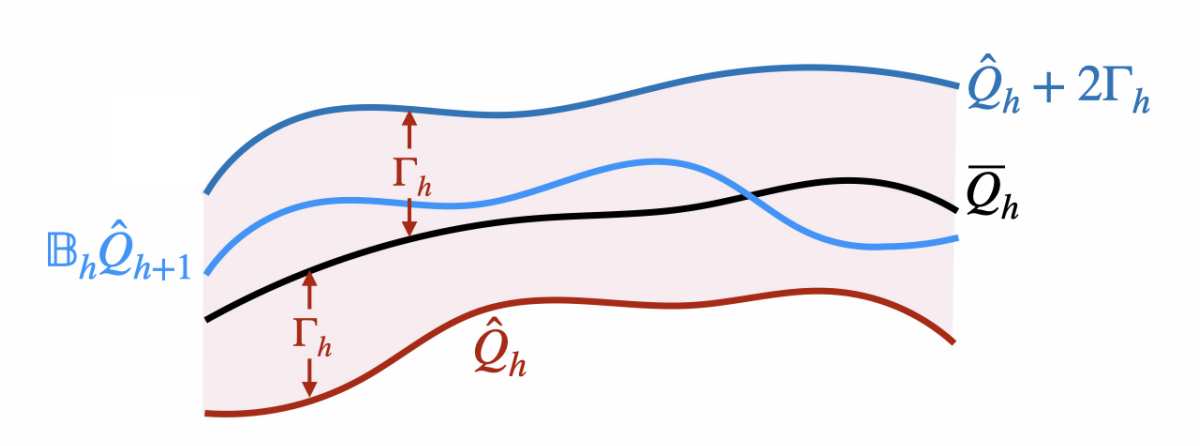
- Construct pessimistic value function

$$\widehat{Q}_h(x,a) = \underbrace{\overline{Q}_h(x,a)}_{\text{VI}} \underbrace{-\Gamma_h(x,a)}_{\text{penalty}}$$

- Optimize: $\widehat{\pi}_h(x) = \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$.

Why Pessimism Helps?

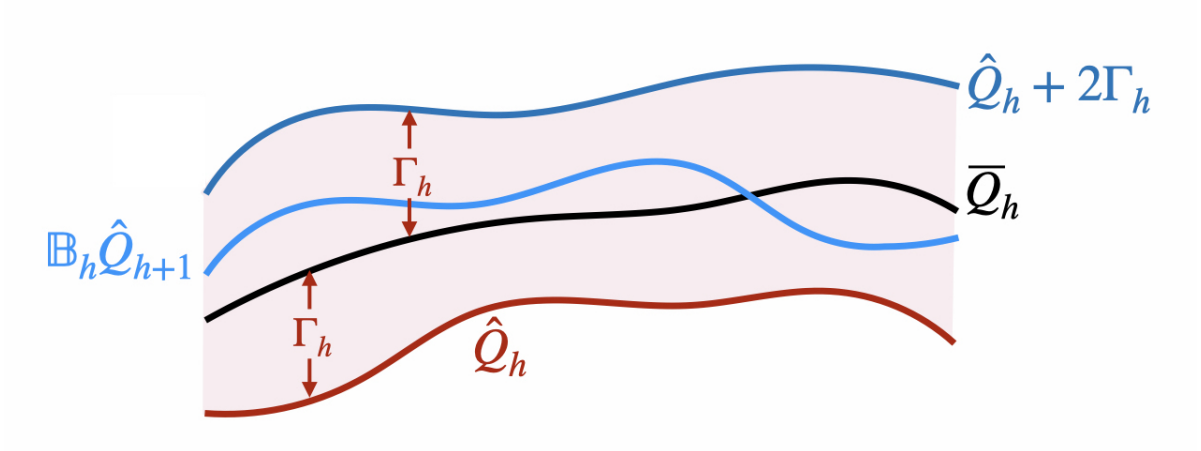
Boundedness of Evaluation Error[1]



[1] Adapted from Lemma 5.1 in (JYW'20)

Why Pessimism Helps?

Boundedness of Evaluation Error[1]



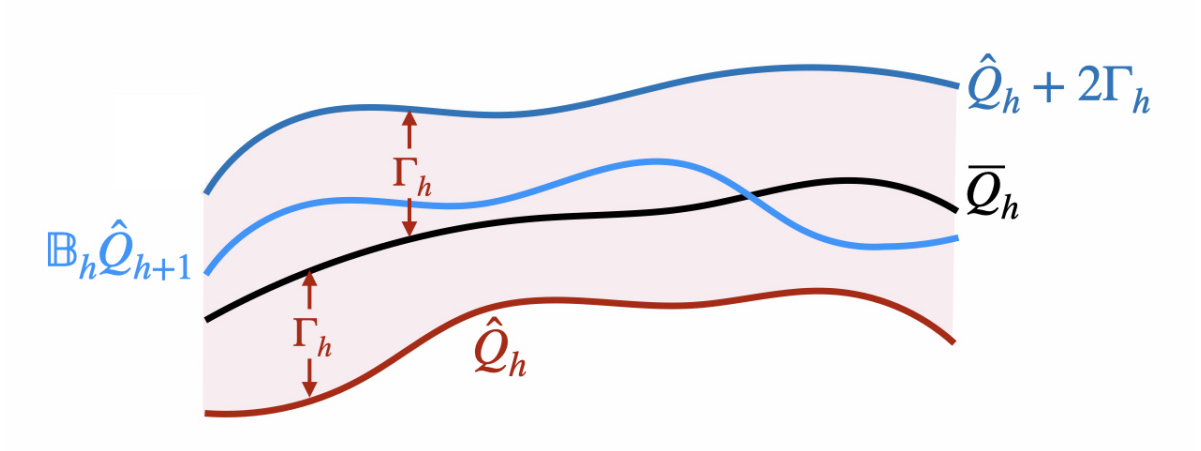
- Define the model evaluation error

$$\iota_h(x,a) = (\mathbb{B}_h\widehat{Q}_{h+1})(x,a) - \widehat{Q}_h(x,a).$$

[1] Adapted from Lemma 5.1 in (JYW'20)

Why Pessimism Helps?

Boundedness of Evaluation Error[1]



- Define the model evaluation error

$$\iota_h(x,a) = (\mathbb{B}_h\widehat{Q}_{h+1})(x,a) - \widehat{Q}_h(x,a).$$

- By pessimistic construction, the model evaluation error satisfy

$$\begin{aligned} & \mathbb{B}_h\widehat{Q}_{h+1} \in \left[\widehat{Q}_h, \widehat{Q}_h + 2\Gamma_h\right] \\ \Rightarrow \quad & 0 \le \iota_h(s_h,a_h) \le 2\Gamma_h(s_h,a_h). \end{aligned}$$

[1] Adapted from Lemma 5.1 in (JYW'20)

Why Pessimism Helps?

Decomposition of Suboptimality[2]

- We have the decomposition of suboptimality

$$\mathsf{SubOpt}(\widehat{\pi}; x) = \underbrace{-\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\iota_h(s_h, a_h) \,\middle|\, s_1 = x\right]}_{\textbf{(i): Spurious Correlation}} + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\iota_h(s_h, a_h) \,\middle|\, s_1 = x\right]}_{\text{(ii): Intrinsic Uncertainty}}$$

$$+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\langle \widehat{Q}_h(s_h, \cdot), \pi_h^\star(\cdot \,|\, s_h) - \widehat{\pi}_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\middle|\, s_1 = x\right]}_{\text{(iii): Optimization Error}}.$$

[2] Adapted from Lemma 3.1 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [3]

- By pessimistic construction, the model evaluation error satisfy

$$0 \leq \iota_h(s_h, a_h) \leq 2\Gamma_h(s_h, a_h).$$

- (i) spurious correlation is always non-positive

$$-\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\iota_h(s_h, a_h) \,\middle|\, s_1 = x\right] \leq 0.$$

[3] Adapted from Theorem 4.2 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [3]

- (i) spurious correlation is always non-positive
- Greedy policy ensures (iii) optimization error is always non-positive

$$\sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\langle \widehat{Q}_h(s_h, \cdot), \pi_h^*(\cdot \,|\, s_h) - \widehat{\pi}_h(\cdot \,|\, s_h) \rangle_{\mathcal{A}} \,\middle|\, s_1 = x \right] \leq 0.$$

[3] Adapted from Theorem 4.2 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [3]

- (i) spurious correlation is always non-positive
- Greedy policy ensures (iii) optimization error is always non-positive
- A clean suboptimality bound

$$\begin{aligned}\mathsf{SubOpt}(\widehat{\pi}; x) &\le \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_h(s_h, a_h) \,\middle|\, s_1 = x\right] \\ &\le 2\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\Gamma_h(s_h, a_h) \,\middle|\, s_1 = x\right]\end{aligned}$$

[3] Adapted from Theorem 4.2 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [3]

- (i) spurious correlation is always non-positive
- Greedy policy ensures (iii) optimization error is always non-positive
- A clean suboptimality bound

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\Gamma_h(s_h, a_h) \,\middle|\, s_1 = x \right]$$

  - Only depends on the trajectory of π^*
  - Pessimism eliminates spurious correlation.

[3] Adapted from Theorem 4.2 in (JYW'20)

Why Pessimism Helps?

Suboptimality Upper Bound [3]

- (i) spurious correlation is always non-positive
- Greedy policy ensures (iii) optimization error is always non-positive
- A clean suboptimality bound

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le 2 \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[\Gamma_h(s_h, a_h) \,\middle|\, s_1 = x \right]$$

  - Only depends on the trajectory of π^*
  - Pessimism eliminates spurious correlation.

Question

How to construct the uncertainty quantifier Γ_h?

[3] Adapted from Theorem 4.2 in (JYW'20)

Instantiation of PEVI

Warm-up: i.i.d. Tabular Case

- Assume $\mathcal{D}$ consists of K i.i.d. trajectories from behavior policy π^b.
- Assume concentrability coefficient

$$\sup_{x,a,h} \frac{\nu_h^\star(x,a)}{\nu_h^b(x,a)} \leq \kappa^\star.$$

Instantiation of PEVI

Warm-up: i.i.d. Tabular Case

- Assume $\mathcal{D}$ consists of K i.i.d. trajectories from behavior policy π^b.
- Assume concentrability coefficient

$$\sup_{x,a,h} \frac{\nu_h^\star(x,a)}{\nu_h^b(x,a)} \leq \kappa^\star.$$

- Uncertainty quantifier

$$\Gamma_h(x,a) \;\propto\; N_h(x,a)^{-1/2}$$

Instantiation of PEVI

Warm-up: i.i.d. Tabular Case

- Assume $\mathcal{D}$ consists of K i.i.d. trajectories from behavior policy π^b.
- Assume concentrability coefficient

$$\sup_{x,a,h} \frac{\nu_h^\star(x,a)}{\nu_h^b(x,a)} \leq \kappa^\star.$$

- Uncertainty quantifier

$$\Gamma_h(x,a) \;\propto\; N_h(x,a)^{-1/2}$$

- The PEVI algorithm gives

$$\mathsf{SubOpt}(\widehat{\pi}; x) \leq \sqrt{S^2 A} \cdot H^2 \sqrt{\kappa^\star / K}.$$

Instantiation of PEVI

Linear MDP

Definition (Linear MDP)

We say an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown (signed) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that

$$\mathbb{P}_h(x' \,|\, x, a) = \langle \phi(x, a), \mu_h(x') \rangle,$$
$$\mathbb{E}\big[r_h(s_h, a_h) \,\big|\, s_h = x, a_h = a\big] = \langle \phi(x, a), \theta_h \rangle$$

for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$. Here we assume $\|\phi(x, a)\| \le 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \le \sqrt{d}$ at each step $h \in [H]$, where $\|\mu_h(\mathcal{S})\| = \int_{\mathcal{S}} \|\mu_h(x)\| \, \mathrm{d}x$.

- Linearity of Bellman update: $\mathbb{B}_h \widehat{Q}_{h+1} = \phi^\top \widehat{\theta}_h$ for some $\widehat{\theta}_h \in \mathbb{R}^d$.
- Linear function approximation $\mathcal{F} = \{f_\theta(x, a) = \phi(x, a)^\top \theta, \ \theta \in \mathbb{R}^d\}$.

Instantiation of PEVI

Linear MDP

Algorithm: PEVI for Linear MDP

- Estimate: $\overline{Q}_h(x,a) = \phi(x,a)^\top \widehat{\theta}_h$ via ridge regression.
- Uncertainty quantification

$$\Gamma_h(x,a) \asymp dH \cdot \left(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\right)^{1/2},$$

  where Λ_h is the augmented sample covariance matrix of $\phi(s_h, a_h)$.

- Pessimistic value function

$$\widehat{Q}_h(x,a) = \phi(x,a)^\top \widehat{\theta}_h - c \cdot dH \cdot \left(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\right)^{1/2}$$

- Optimize: $\widehat{\pi}_h(x) = \operatorname{argmax}_{a \in \mathcal{A}} \widehat{Q}_h(x,a)$.

Instantiation of PEVI - Linear MDP

Compliance Assumption

Assumption: Compliance

Let $\mathbb{P}_{\mathcal{D}}$ be the joint distribution of the dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$. We say $\mathcal{D}$ is compliant with an MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ if

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}}\big(r_h^\tau = r', x_{h+1}^\tau = x' \,\big|\, \{(x_h^j, a_h^j)\}_{j=1}^{\tau}, \{(r_h^j, x_{h+1}^j)\}_{j=1}^{\tau-1}\big)\\ &= \mathbb{P}\big(r_h = r', s_{h+1} = x' \,\big|\, s_h = x_h^\tau, a_h = a_h^\tau\big)\end{aligned}$$

for all $r' \in [0,1]$, $x' \in \mathcal{S}$, $h \in [H]$, $\tau \in [K]$. Here $\mathbb{P}$ is taken with respect to the underlying MDP.

- Only require that $\mathcal{D}$ evolves according to the MDP.
- Minimal assumptions on actions a_h^τ: allow for arbitrarily collected data.
  - i.i.d. trajectories from a behavior policy ✓
  - sequentially adjusted actions $a_h^\tau \in \sigma(\{x_{h+1}^j, r_h^j\}_{j<\tau})$ ✓

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound

Theorem 4.4 (JYW'20)

If $\mathcal{D}$ is compliant with the underlying MDP, then w.h.p,

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\right)^{1/2} \,\middle|\, s_1 = x\right].$$

up to logarithm factors of d, H, K.

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound

Theorem 4.4 (JYW'20)

If $\mathcal{D}$ is compliant with the underlying MDP, then w.h.p,

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\right)^{1/2} \,\middle|\, s_1 = x\right].$$

up to logarithm factors of d, H, K.

- Minimal-assumption guarantee: only require compliance of $\mathcal{D}$.

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound

Theorem 4.4 (JYW'20)

If $\mathcal{D}$ is compliant with the underlying MDP, then w.h.p,

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\right)^{1/2} \middle| s_1 = x\right].$$

up to logarithm factors of d, H, K.

- Minimal-assumption guarantee: only require compliance of $\mathcal{D}$.
- Oracle property: only depends on how well $\pi^\star$ is covered - no requirement on coverage of all trajectories.

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound

Theorem 4.4 (JYW'20)

If $\mathcal{D}$ is compliant with the underlying MDP, then w.h.p,

$$\mathsf{SubOpt}(\widehat{\pi}; x) \le c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\right)^{1/2} \,\middle|\, s_1 = x\right].$$

up to logarithm factors of d, H, K.

- Minimal-assumption guarantee: only require compliance of $\mathcal{D}$.
- Oracle property: only depends on how well $\pi^\star$ is covered - no requirement on coverage of all trajectories.
- Data-dependent upper bound: (offline) data is what it is.

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound: Special Cases

- Well-explored case: $\approx$ uniformly good coverage of all policies.
- The suboptimality achieves $\sqrt{1/K}$ rate if $\mathcal{D}$ consist of K i.i.d. trajectories from behavior policy $\bar{\pi}$ and

$$\lambda_{\min}\left(\mathbb{E}_{\bar{\pi}}[\phi(s_h, a_h)\phi(s_h, a_h)^\top]\right) \geq c \quad \text{for some } c > 0.$$

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound: Special Cases

- Essentially-explored case: $\approx$ good coverage of optimal policy.
- The suboptimality achieves $\sqrt{1/K}$ rate if

$$\Lambda_h \succeq \mathbf{I} + c \cdot K \cdot \mathbb{E}_{\pi^\star}\left[\phi(s_h, a_h)\phi(s_h, a_h)^\top \,\middle|\, s_1 = x\right].$$

Instantiation of PEVI - Linear MDP

Suboptimality Upper Bound: Special Cases

- Essentially-explored case: $\approx$ good coverage of optimal policy.
- The suboptimality achieves $\sqrt{1/K}$ rate if

$$\Lambda_h \succeq \mathbf{I} + c \cdot K \cdot \mathbb{E}_{\pi^\star}\left[\phi(s_h, a_h)\phi(s_h, a_h)^\top \,\middle|\, s_1 = x\right].$$

Question

Is coverage of optimal $\pi^\star$ the essential information in $\mathcal{D}$?

Is Pessimism Efficient?

- Minimax Lower Bounds for Linear MDP

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality in Linear MDP

- Upper bound: pessimistic policy $\widehat{\pi}$ and compliant $\mathcal{D} \sim \mathcal{M}$,

$$\mathsf{SubOpt}(\mathcal{M}, \widehat{\pi}; x) \leq c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\right)^{1/2} \middle| s_1 = x\right].$$

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality in Linear MDP

- Upper bound: pessimistic policy $\widehat{\pi}$ and compliant $\mathcal{D} \sim \mathcal{M}$,

$$\mathsf{SubOpt}\big(\mathcal{M}, \widehat{\pi}; x\big) \leq c \cdot dH \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \Big| s_1 = x\Big].$$

- Lower bound: for any offline learning algorithm $\texttt{Algo}(\cdot)$,

$$\sup_{\mathcal{M}, \mathcal{D}} \mathbb{E}_{\mathcal{D}}\left[\frac{\mathsf{SubOpt}\big(\mathcal{M}, \texttt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \Big| s_1 = x\Big]}\right] \geq c.$$

Minimax Optimality of Pessimism: Linear MDP

- Answer: Coverage of optimal $\pi^\star$ is the essential information in $\mathcal{D}$.
- Pessimism is (nearly) minimax optimal in linear setting.

Minimax Optimality in Linear MDP

- Lower bound: for any offline learning algorithm $\texttt{Algo}(\cdot)$,

$$\sup_{\mathcal{M},\mathcal{D}} \mathbb{E}_{\mathcal{D}}\left[\frac{\mathsf{SubOpt}\big(\mathcal{M}, \texttt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left(\phi(s_h,a_h)^\top \Lambda_h^{-1} \phi(s_h,a_h)\right)^{1/2} \,\middle|\, s_1 = x\right]}\right] \geq c.$$

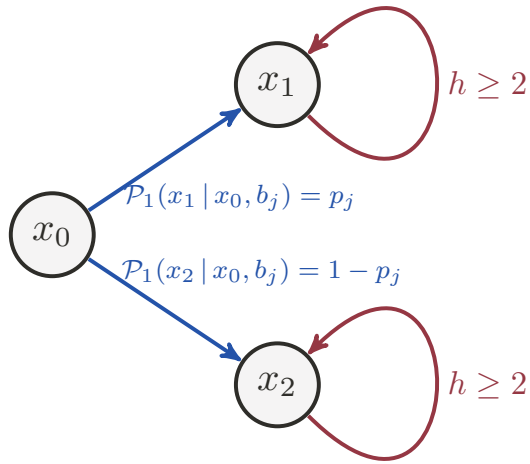
  - Dependence on true MDP $\mathcal{M}$ and its optimal policy $\pi^\star$.
  - Essential Hardness in $\mathcal{D}$: how well (sample covariance) Λ_h covers $\pi^\star$.

Thank you!

Proof of Minimax Lower Bound

Construction of Hard Instance

- A subclass of linear MDP $M(p_1, p_2, p_3)$.



  - Actions $\mathcal{A} = \{b_1, b_2, \ldots, b_A\}$, states $\mathcal{S} = \{x_0, x_1, x_2\}$.
  - Initial state $s_1 \equiv x_0$.
  - Transition at the first step $\mathbb{P}_1(x_1 \mid x_0, b_j) = p_j, \quad p_j = p_3$ for $j \geq 3$.
  - Absorbing states x_1, x_2, $\mathbb{P}_h(x_i \mid x_i, a) = 1$ for $i = 1, 2$ and all $a \in \mathcal{A}$.
  - Deterministic rewards $r_h(x_1, a) = 1$ and $r_h(x_2, a) = 0$ for all $a \in \mathcal{A}$.

- Pre-determined actions in $\mathcal{D}$: take action b_j for n_j times.

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Two instances $\mathcal{M}_1 = M(p^*, p, p)$, $\mathcal{M}_2 = M(p, p^*, p)$, $p < p^*$.
  - Optimal policy for $\mathcal{M}_1$ takes b_1, while optimal policy for $\mathcal{M}_2$ takes b_2.

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Two instances $\mathcal{M}_1 = M(p^*, p, p)$, $\mathcal{M}_2 = M(p, p^*, p)$, $p < p^*$.
  - Optimal policy for $\mathcal{M}_1$ takes b_1, while optimal policy for $\mathcal{M}_2$ takes b_2.
- Suboptimality of any policy π on the two MDPs are

$$\mathsf{SubOpt}\big(\mathcal{M}_\ell, \pi; x_0\big) = (p^* - p)(H-1)\big(1 - \pi_1(b_\ell \,|\, x_0)\big)$$

  - 2-point argument: any policy makes mistake either on $\mathcal{M}_1$ or on $\mathcal{M}_2$.

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Two instances $\mathcal{M}_1 = M(p^*, p, p)$, $\mathcal{M}_2 = M(p, p^*, p)$, $p < p^*$.
  - Optimal policy for $\mathcal{M}_1$ takes b_1, while optimal policy for $\mathcal{M}_2$ takes b_2.
- Suboptimality of any policy π on the two MDPs are

$$\mathsf{SubOpt}\big(\mathcal{M}_\ell, \pi; x_0\big) = (p^* - p)(H-1)\big(1 - \pi_1(b_\ell \,|\, x_0)\big)$$

  - 2-point argument: any policy makes mistake either on $\mathcal{M}_1$ or on $\mathcal{M}_2$.

- Reduction to Testing:

$$\begin{aligned}
&\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell}\Big[\mathsf{SubOpt}\big(\mathcal{M}_\ell, \texttt{Algo}(\mathcal{D}); x_0\big)\Big] \\
&\geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1) \\
&\qquad \times \Big(\mathbb{E}_{\mathcal{D} \sim \mathcal{M}_1}\big[1 - \pi_1(b_1 \,|\, x_0)\big] + \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_2}\big[\pi_1(b_1 \,|\, x_0)\big]\Big) \\
&\geq \frac{\sqrt{n_1 n_2}}{\sqrt{n_1} + \sqrt{n_2}} \cdot (p^* - p) \cdot (H-1)\Big(1 - \mathrm{TV}(\mathbb{P}_{\mathcal{D} \sim \mathcal{M}_1}, \mathbb{P}_{\mathcal{D} \sim \mathcal{M}_2})\Big)
\end{aligned}$$

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Reduction to Testing:

$$\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell} \Big[\mathsf{SubOpt}\big(\mathcal{M}_\ell, \mathtt{Algo}(\mathcal{D}); x_0\big) \Big] \geq c \cdot (H-1),$$

  with a careful choice of n_1, n_2 and p, p^*.

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Reduction to Testing:

$$\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell} \Big[\mathsf{SubOpt}\big(\mathcal{M}_\ell, \mathtt{Algo}(\mathcal{D}); x_0\big) \Big] \geq c \cdot (H-1),$$

  with a careful choice of n_1, n_2 and p, p^*.

- At the same time, w.h.p.,

$$\sum_{h=1}^{H} \mathbb{E}_{\pi^*, \ell, \mathcal{M}_\ell} \Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0 \Big] \approx (H-1)/\sqrt{n_\ell}.$$

Proof of Minimax Lower Bound

Modified LeCam 2-Point Argument

- Reduction to Testing:

$$\max_{\ell \in \{1,2\}} \sqrt{n_\ell} \cdot \mathbb{E}_{\mathcal{D} \sim \mathcal{M}_\ell} \Big[\mathsf{SubOpt}\big(\mathcal{M}_\ell, \mathtt{Algo}(\mathcal{D}); x_0\big) \Big] \geq c \cdot (H-1),$$

  with a careful choice of n_1, n_2 and p, p^*.

- At the same time, w.h.p.,

$$\sum_{h=1}^{H} \mathbb{E}_{\pi^*, \ell, \mathcal{M}_\ell} \Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x_0 \Big] \approx (H-1)/\sqrt{n_\ell}.$$

- Final lower bound

$$\sup_{\mathcal{M}, \mathcal{D}} \mathbb{E}_{\mathcal{D}} \left[\frac{\mathsf{SubOpt}\big(\mathcal{M}, \mathtt{Algo}(\mathcal{D}); x\big)}{\sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \Big[\big(\phi(s_h, a_h)^\top \Lambda_h^{-1} \phi(s_h, a_h)\big)^{1/2} \,\Big|\, s_1 = x \Big]} \right] \geq c.$$

Proof of Upper Bound for Linear MDP

Formula of Uncertainty Quantifier

- Bellman update $\mathbb{B}_h \widehat{Q}_{h+1}(x,a) = \phi(x,a)^\top w_h$ for some $w_h \in \mathbb{R}^d$.

Proof of Upper Bound for Linear MDP

Formula of Uncertainty Quantifier

- Bellman update $\mathbb{B}_h \widehat{Q}_{h+1}(x,a) = \phi(x,a)^\top w_h$ for some $w_h \in \mathbb{R}^d$.
- Ridge estimator $\bar{Q}_h(x,a) = \phi(x,a)^\top \widehat{w}_h$, with

$$\widehat{w}_h = \Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\Big),$$

$$\text{where} \quad \widehat{V}_{h+1}(x) = \max_{a \in \mathcal{A}} \widehat{Q}_{h+1}(x,a),$$

$$\Lambda_h = \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I},$$

Proof of Upper Bound for Linear MDP

Formula of Uncertainty Quantifier

- Bellman update $\mathbb{B}_h \widehat{Q}_{h+1}(x,a) = \phi(x,a)^\top w_h$ for some $w_h \in \mathbb{R}^d$.
- Ridge estimator $\bar{Q}_h(x,a) = \phi(x,a)^\top \widehat{w}_h$, with

$$\widehat{w}_h = \Lambda_h^{-1}\Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)\big)\Big),$$

$$\text{where} \quad \widehat{V}_{h+1}(x) = \max_{a \in \mathcal{A}} \widehat{Q}_{h+1}(x,a),$$

$$\Lambda_h = \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot \mathbf{I},$$

- Uncertainty quantifier chosen as

$$\Gamma_h(x,a) = \beta \cdot \big(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\big)^{1/2},$$

where the constant β is to be specified.

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Validity of Γ_h**: w.h.p. for all (x, a) and all $h \in [H]$,

$$\left|\mathbb{B}_h \widehat{Q}_{h+1}(x,a) - \bar{Q}_h(x,a)\right| \le \Gamma_h(x,a) = \beta \cdot \left(\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)\right)^{1/2}$$

- The difference is decomposed into

$$\begin{aligned}
&(\mathbb{B}_h \widehat{Q}_{h+1})(x,a) - \bar{Q}_h(x,a) = \phi(x,a)^\top (w_h - \widehat{w}_h) \\
&= \underbrace{\phi(x,a)^\top w_h - \phi(x,a)^\top \Lambda_h^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (\mathbb{B}_h \widehat{Q}_{h+1})(x_h^\tau, a_h^\tau) \Big)}_{\text{(i)}} \\
&\quad \underbrace{- \phi(x,a)^\top \Lambda_h^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \big(r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau) - (\mathbb{B}_h \widehat{Q}_{h+1})(x_h^\tau, a_h^\tau)\big) \Big)}_{\text{(ii)}} .
\end{aligned}$$

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Boundedness of w_h**: since $\widehat{Q}_{h+1} \in [0, H-h]$, it holds that

$$\|\widehat{w}_h\| \leq H\sqrt{Kd/\lambda}.$$

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Boundedness of** w_h: since $\widehat{Q}_{h+1} \in [0, H-h]$, it holds that

$$\|\widehat{w}_h\| \leq H\sqrt{Kd/\lambda}.$$

- The first term bounded as

$$\big|(\mathsf{i})\big| = \lambda \cdot \big|\phi(x,a)^\top \Lambda_h^{-1} w_h\big| \leq H\sqrt{d\lambda} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}.$$

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Boundedness of** w_h: since $\widehat{Q}_{h+1} \in [0, H-h]$, it holds that

$$\|\widehat{w}_h\| \le H\sqrt{Kd/\lambda}.$$

- The first term bounded as

$$\big|(\mathsf{i})\big| = \lambda \cdot \big|\phi(x,a)^\top \Lambda_h^{-1} w_h\big| \le H\sqrt{d\lambda} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}.$$

- The second term bounded as

$$\big|(\mathsf{ii})\big| \le \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\widehat{V}_{h+1}) \Big\|_{\Lambda_h^{-1}} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}$$

where $\epsilon_h^\tau(V) = r_h^\tau + V(x_{h+1}^\tau) - \mathbb{E}\big[r_h(s_h,a_h) + V(s_{h+1}) \,\big|\, s_h = x_h^\tau, a_h = a_h^\tau\big]$.

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Uniform concentration**:

$$\big|\text{(ii)}\big| \le \sup_{V_{h+1}\in\mathcal{V}_{h+1}} \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\cdot \epsilon_h^\tau(V_{h+1}) \Big\|_{\Lambda_h^{-1}} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}.$$

- Supremum over function class $\mathcal{V}_{h+1}$ with the form

$$V_{h+1}(x) = \max_{a\in\mathcal{A}} \Big\{ \min\big\{ \phi(x,a)^\top \theta - \beta \cdot \sqrt{\phi(x,a)^\top \Sigma^{-1} \phi(x,a)}, H-h \big\}^+ \Big\},$$

for appropriately bounded $\theta \in \mathbb{R}^d$, $\beta \in \mathbb{R}$ and $\Sigma \in \mathbb{R}^{d\times d}$.

Proof of Upper Bound for Linear MDP

Validity of Uncertainty Quantifier

- **Uniform concentration**:

$$\big|\text{(ii)}\big| \le \sup_{V_{h+1}\in\mathcal{V}_{h+1}} \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\cdot \epsilon_h^\tau(V_{h+1})\Big\|_{\Lambda_h^{-1}} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)}.$$

- Supremum over function class $\mathcal{V}_{h+1}$ with the form

$$V_{h+1}(x) = \max_{a\in\mathcal{A}}\Big\{\min\big\{\phi(x,a)^\top\theta - \beta\cdot\sqrt{\phi(x,a)^\top\Sigma^{-1}\phi(x,a)}, H-h\big\}^+\Big\},$$

  for appropriately bounded $\theta\in\mathbb{R}^d$, $\beta\in\mathbb{R}$ and $\Sigma\in\mathbb{R}^{d\times d}$.

- The second term is bounded as

$$\big|\text{(ii)}\big| \le \beta/2\cdot\sqrt{\phi(x,a)^\top\Lambda_h^{-1}\phi(x,a)}, \quad \beta = c\cdot dH\cdot \mathsf{PolyLog}(d,H,K)$$

  - Concentration of self-normalized process for a single V_{h+1}. (Only compliance of $\mathcal{D}$ is needed for the concentration.)
  - ε-covering of linear function class $\mathcal{V}_{h+1}$ for uniform concentration.