Bellman-consistent Pessimism for Offline Reinforcement Learning

Tengyang Xie[1], Ching-An Cheng[2], Nan Jiang[1], Paul Mineiro[2], Alekh Agarwal[3]

1 

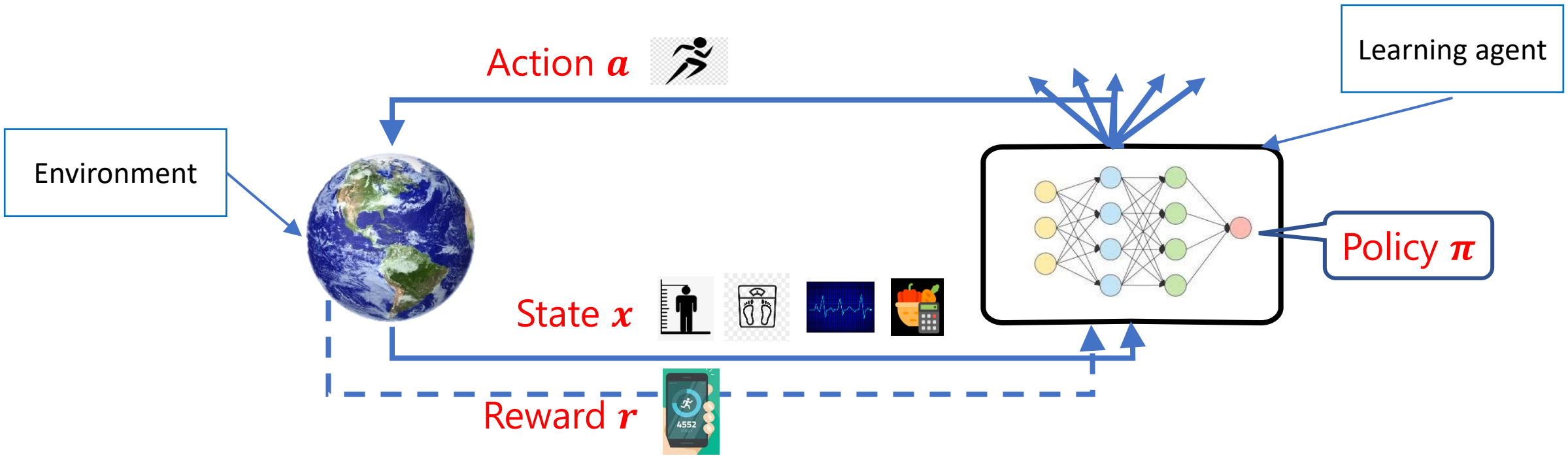
2 

3 

Agenda

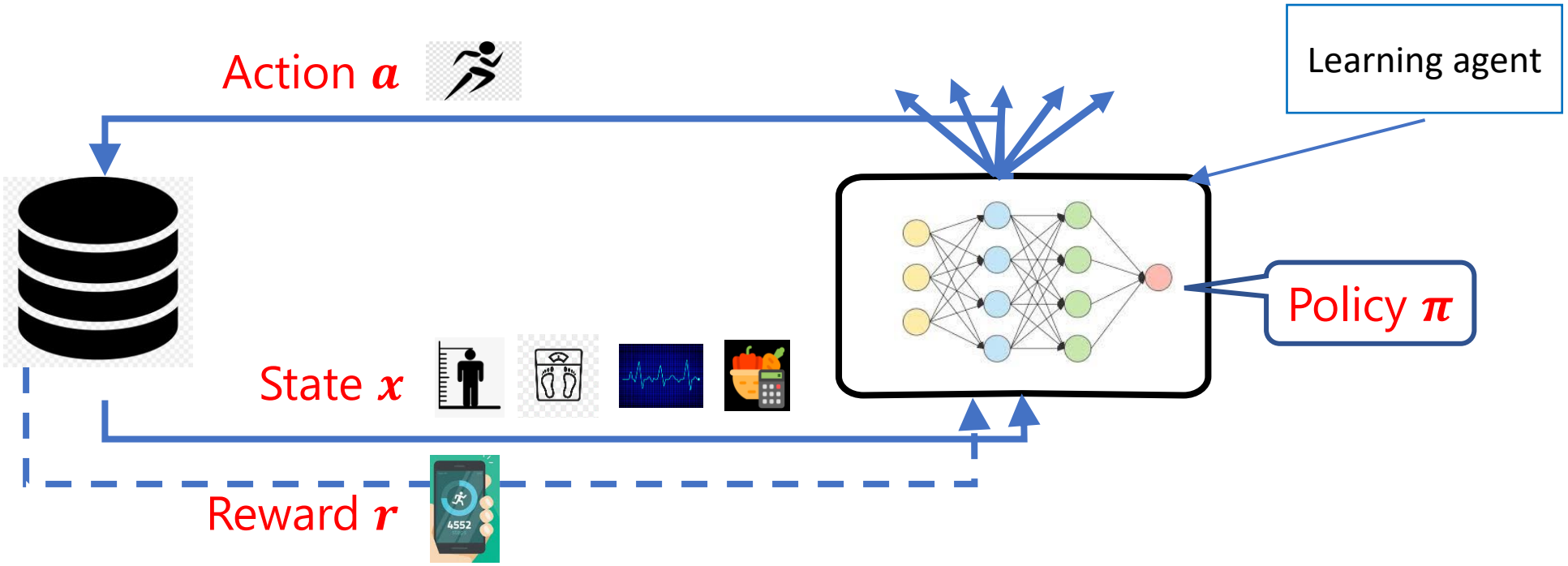
- **Problem Setup**
- Bellman-consistent Pessimism
- Practical Algorithm
- Related Works and Summary

Reinforcement Learning



Goal: **Learn a policy π** to maximize the cumulative rewards

Offline Reinforcement Learning



- Learning from offline data only, no interaction access.
- No *coverage* assumption on offline data.

Offline RL: Notations and Basics

- Infinite horizon MDPs with discount factor γ
- Bellman operator: $\mathcal{T}^{\pi} f(s,a) \coloneqq R(s,a) + \gamma \mathbb{E}_{s'|s,a}[f(s',\pi)]$
- Value function:
  - $Q^{\pi} \longleftarrow$ Fixed point of $\mathcal{T}^{\pi}$
  - $J(\pi) \coloneqq \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t r_t] = Q^{\pi}(s_0,\pi)$
- Discounted state-action occupancy: $d_{\pi}(s,a) \coloneqq (1-\gamma)\mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}(s_t = s, a_t = a)]$
- Offline setting: offline data $\mathcal{D} \sim \mu$ ($(s,a) \sim \mu, r \sim R, s' \sim P$), no interaction with environment.

No *coverage* assumption on offline data.

Problem Setup

- What we have?

  1. An offline dataset $\mathcal{D} \sim \mu$ (no interaction w/ env.)

  2. Policy class Π and function class $\mathcal{F}$

- Assumptions:

  - **Realizability**: (approximately) $Q^\pi \in \mathcal{F}, \forall \pi \in \Pi$

  - **Completeness**: (approximately) $\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \forall \pi \in \Pi$

$$\inf_{f \in \mathcal{F}} \sup_{\text{admissible } \nu} \|f - \mathcal{T}^\pi f\|_{2,\nu}^2 \leq \varepsilon_{\mathcal{F}}, \forall \pi \in \Pi$$

$$\sup_{f \in \mathcal{F}} \inf_{f' \in \mathcal{F}} \|f' - \mathcal{T}^\pi f\|_{2,\mu}^2 \leq \varepsilon_{\mathcal{F},\mathcal{F}}, \forall \pi \in \Pi$$

Desired Guarantee in Offline RL

- When single-policy concentrability is assumed for an arbitrary π

$$C_\pi := \max_{f, f' \in \mathcal{F}} \frac{\|f - f'\|_{2,d_\pi}^2}{\|f - f'\|_{2,\mu}^2} \leq \|d_\pi / \mu\|_\infty$$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{C_\pi \log(|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{\sqrt{C_\pi (\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}})}}{1-\gamma}$$

Desired Guarantee in Offline RL

- When single-policy concentrability is assumed for an arbitrary π

$$C_\pi := \max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,d_\pi}^2}{\|f - f'\|_{2,\mu}^2} \leq \|d_\pi/\mu\|_\infty$$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{C_\pi \log(|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{\sqrt{C_\pi}(\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}})}{1-\gamma}$$

"uniform" concentrability ($\max_\pi C_\pi$) [e.g., Xie and Jiang 19] ⟶ single-policy concentrability (C_π)

[Xie and Jiang 19] **Q* Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison** UAI'20

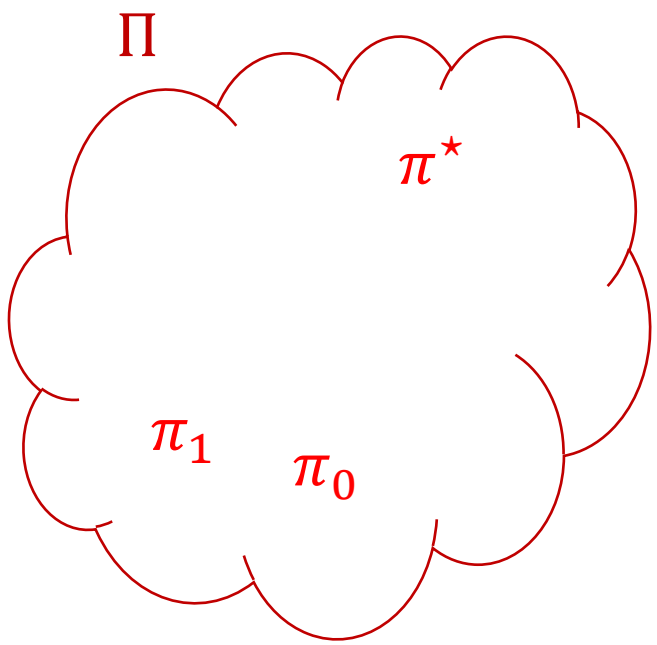
Agenda

- Problem Setup
- **Bellman-consistent Pessimism**
- Practical Algorithm
- Related Works and Summary

Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Goal of RL:

$$\pi^\star = \operatorname*{argmax}_{\pi \in \Pi} J(\pi) = \operatorname*{argmax}_{\pi \in \Pi} Q^\pi(s_0, \pi)$$

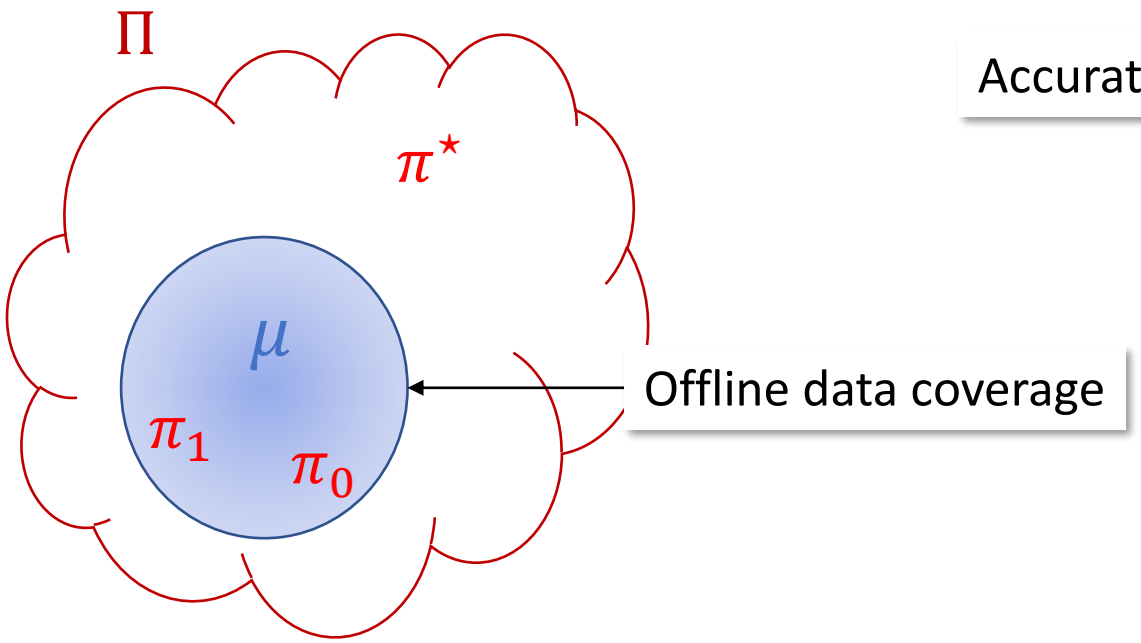


Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Goal of RL:

$$\pi^\star = \operatorname*{argmax}_{\pi\in\Pi} J(\pi) = \operatorname*{argmax}_{\pi\in\Pi} Q^\pi(s_0,\pi)$$

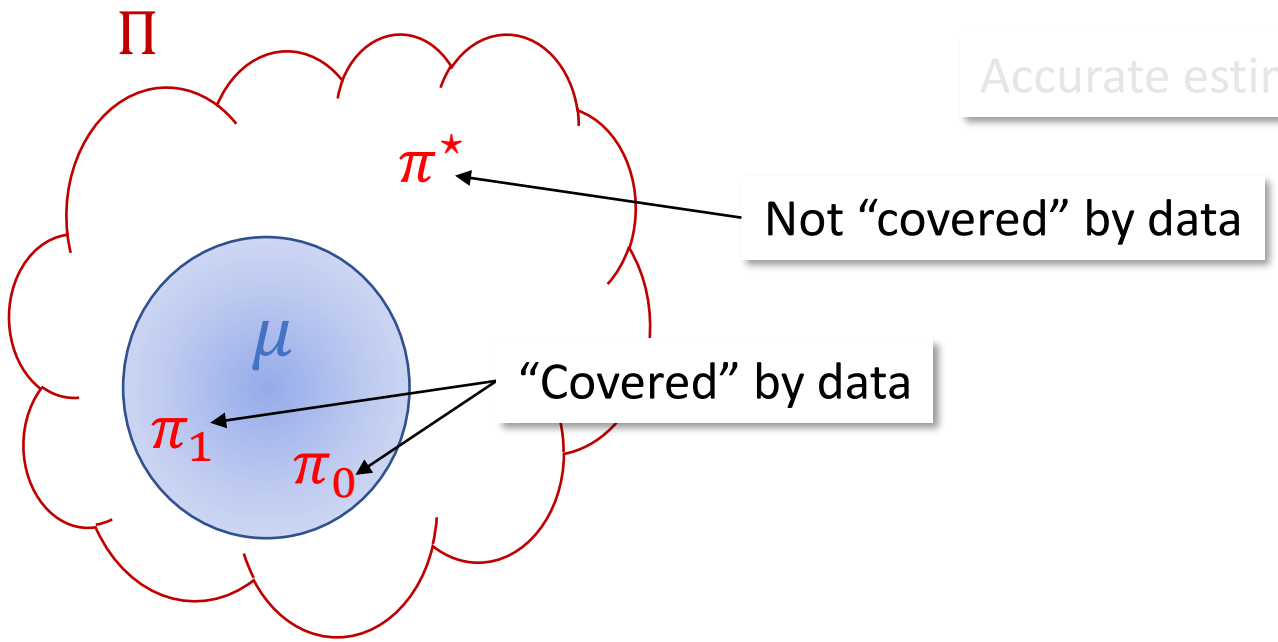
Accurate estimation NOT always available in **offline** setting



Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Goal of RL:

$$\pi^\star = \operatorname*{argmax}_{\pi\in\Pi} J(\pi) = \operatorname*{argmax}_{\pi\in\Pi} Q^\pi(s_0,\pi)$$



Accurate estimation NOT always available in **offline** setting

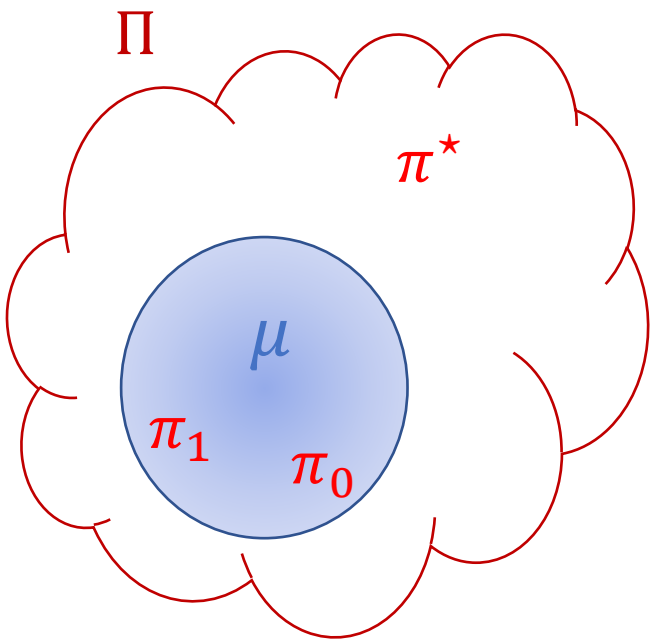
Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Our proposal for **offline** RL:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \hat{Q}^{\pi}_{\mathrm{pes}}(s_0, \pi)$$

Policy's best pessimistic evaluation based on what we have

Given $\mathcal{D}$ and $\mathcal{F}$, what is the worst possible performance of π?



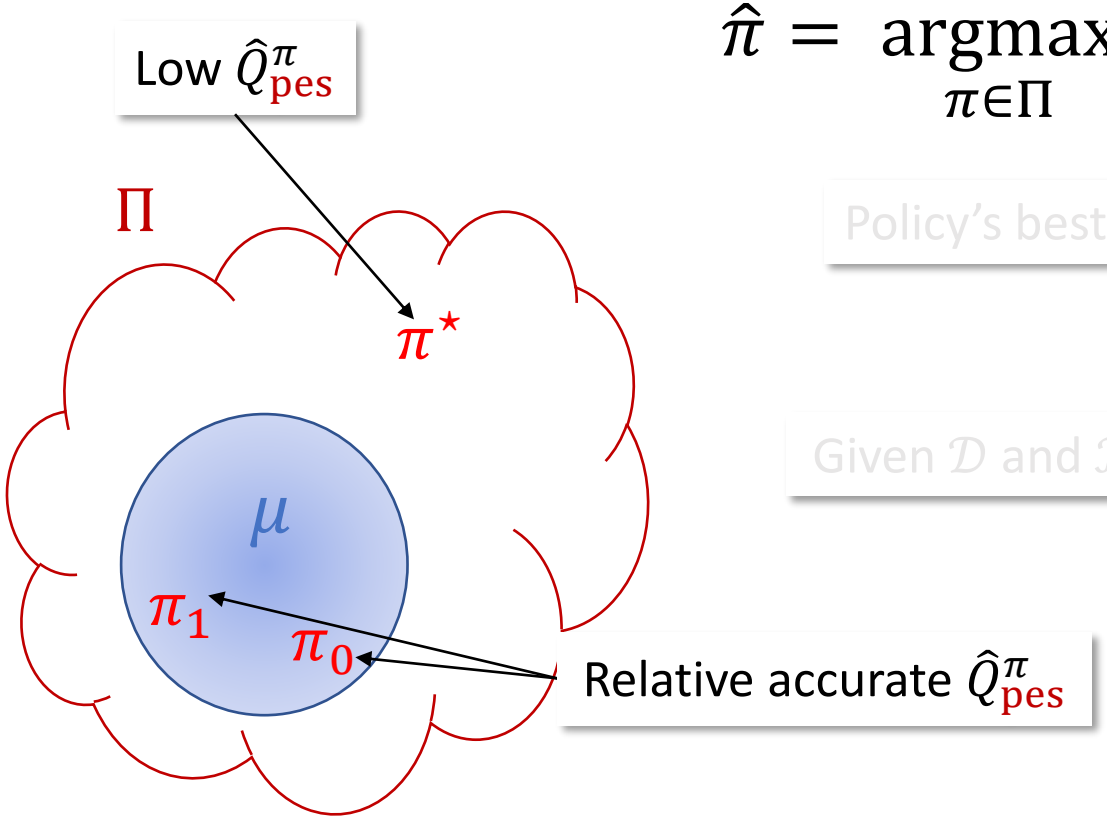
Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Our proposal for **offline** RL:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \hat{Q}^{\pi}_{\mathrm{pes}}(s_0, \pi)$$

Policy's best pessimistic evaluation based on what we have

Given $\mathcal{D}$ and $\mathcal{F}$, what is the worst possible performance of π?



Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Information-theoretic algorithm for Bellman-consistent Pessimism:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi),$$

$$\mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \leq \varepsilon\}$$

All plausible Q-functions based that is consistent with offline data (including Q^{π})

Bellman-consistent Pessimism ---A General Theoretical Framework for Offline RL

- Information-theoretic algorithm for Bellman-consistent Pessimism:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi),$$

$$\mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \leq \varepsilon\}$$

$:= \hat{Q}^{\pi}_{\text{pes}}$

All possible Q-functions based on our offline data (including Q^{π})

$$\hat{Q}^{\pi}_{\text{pes}} = \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) \leq J(\pi)$$

Bellman-consistent Pessimism
---A General Theoretical Framework for Offline RL

- Information-theoretic algorithm for Bellman-consistent Pessimism:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi), \qquad \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \leq \varepsilon\}$$

$$\approx \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a) - r - \gamma f(s', \pi)\right)^2\right] - \min_{f' \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a) - r - \gamma f(s', \pi)\right)^2\right]$$

[Antos et al. 08] **Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path**

Derivation

Most optimistic estimation of $J(\pi)$ in the version space

Most pessimistic estimation of $J(\hat{\pi})$ in the version space

- $J(\pi) - J(\hat{\pi}) \leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\hat{\pi},\varepsilon}} f(s_0, \hat{\pi})$

Derivation

- $J(\pi) - J(\hat{\pi}) \leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\hat{\pi},\varepsilon}} f(s_0, \hat{\pi})$

$$\leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$$

$\coloneqq f_{\pi,\max}(s_0, \pi)$

$\coloneqq f_{\pi,\min}(s_0, \pi)$

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$$

$$\min_{f \in \mathcal{F}_{\hat{\pi},\varepsilon}} f(s_0, \hat{\pi}) \geq \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$$

Derivation

- $J(\pi) - J(\hat{\pi}) \leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\hat{\pi},\varepsilon}} f(s_0, \hat{\pi})$

$$\leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$$

Standard telescoping argument

$$\leq \frac{\mathbb{E}_{d_\pi}[f_{\pi,\max} - \mathcal{T}^\pi f_{\pi,\max}] - \mathbb{E}_{d_\pi}[f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}]}{1-\gamma}$$

Derivation

- $J(\pi) - J(\hat{\pi}) \leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\hat{\pi},\varepsilon}} f(s_0, \hat{\pi})$

$$\leq \max_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi) - \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi)$$

$$\leq \frac{\mathbb{E}_{d_\pi}[f_{\pi,\max} - \mathcal{T}^\pi f_{\pi,\max}] - \mathbb{E}_{d_\pi}[f_{\pi,\min} - \mathcal{T}^\pi f_{\pi,\min}]}{1-\gamma}$$

$= \mathbb{E}_\nu + \mathbb{E}_{d_\pi \setminus \nu}$

Some on-support distribution

Pay $\max_{f,f' \in \mathcal{F}} \frac{\|f-f'\|_{2,\nu}^2}{\|f-f'\|_{2,\mu}^2} \leq \|\nu/\mu\|_\infty^2$ in concentration

$:= \max\{d_\pi(s,a) - \nu(s,a), 0\}$
Controls off-support mass

Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For *any* $\pi \in \Pi$, *<u>any</u>* constant C, and *<u>any</u>* distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$, $\leq \|\nu/\mu\|_\infty^2$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{C \log(|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{\langle d_\pi \setminus \nu, \Delta f_\pi - \mathcal{T}^\pi \Delta f_\pi \rangle}{1-\gamma}$$

Distribution shift between ν and data distribution μ

on-support part (variance)

off-support part (bias)

Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For *any* $\pi \in \Pi$, *any* constant C, and *any* distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{C \log(|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{\langle d_\pi \setminus \nu, \Delta f_\pi - \mathcal{T}^\pi \Delta f_\pi \rangle}{1-\gamma}$$

$\coloneqq f_{\pi,\max} - f_{\pi,\min}$

$\coloneqq \mathcal{T}^\pi f_{\pi,\max} - \mathcal{T}^\pi f_{\pi,\min}$

$\coloneqq \max\{d_\pi(s,a) - \nu(s,a), 0\}$

Distribution shift between ν and data distribution μ

on-support part (variance)

off-support part (bias)

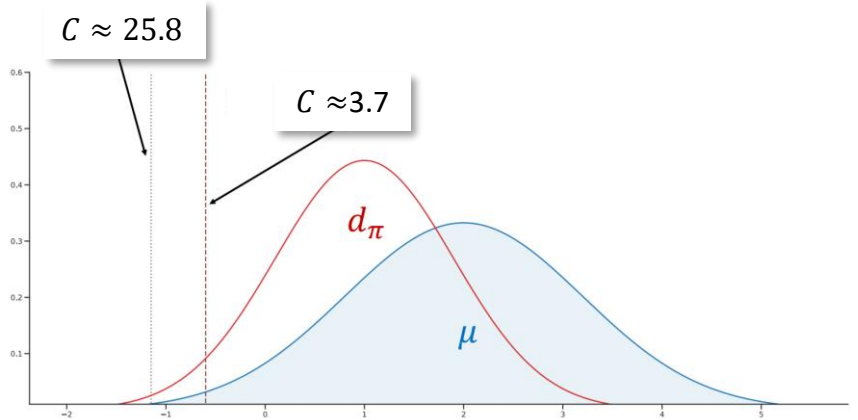
Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For *any* $\pi \in \Pi$, *<u>any</u>* constant C, and *<u>any</u>* distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$

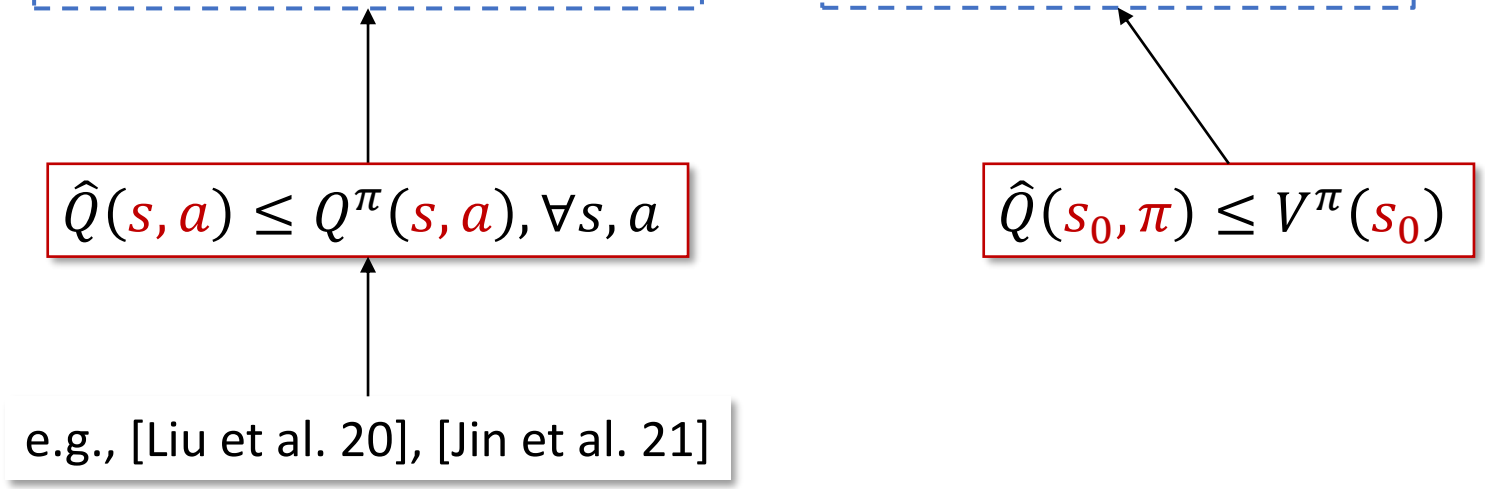
$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma} \sqrt{\frac{C \log(|\mathcal{F}||\Pi|/\delta)}{n}} + \frac{\langle d_\pi \setminus \nu, \Delta f_\pi - \mathcal{T}^\pi \Delta f_\pi \rangle}{1-\gamma}$$

On-support off-support splitting ⟺ “bias-variance trade-off”



Comparison between Pessimism Frameworks

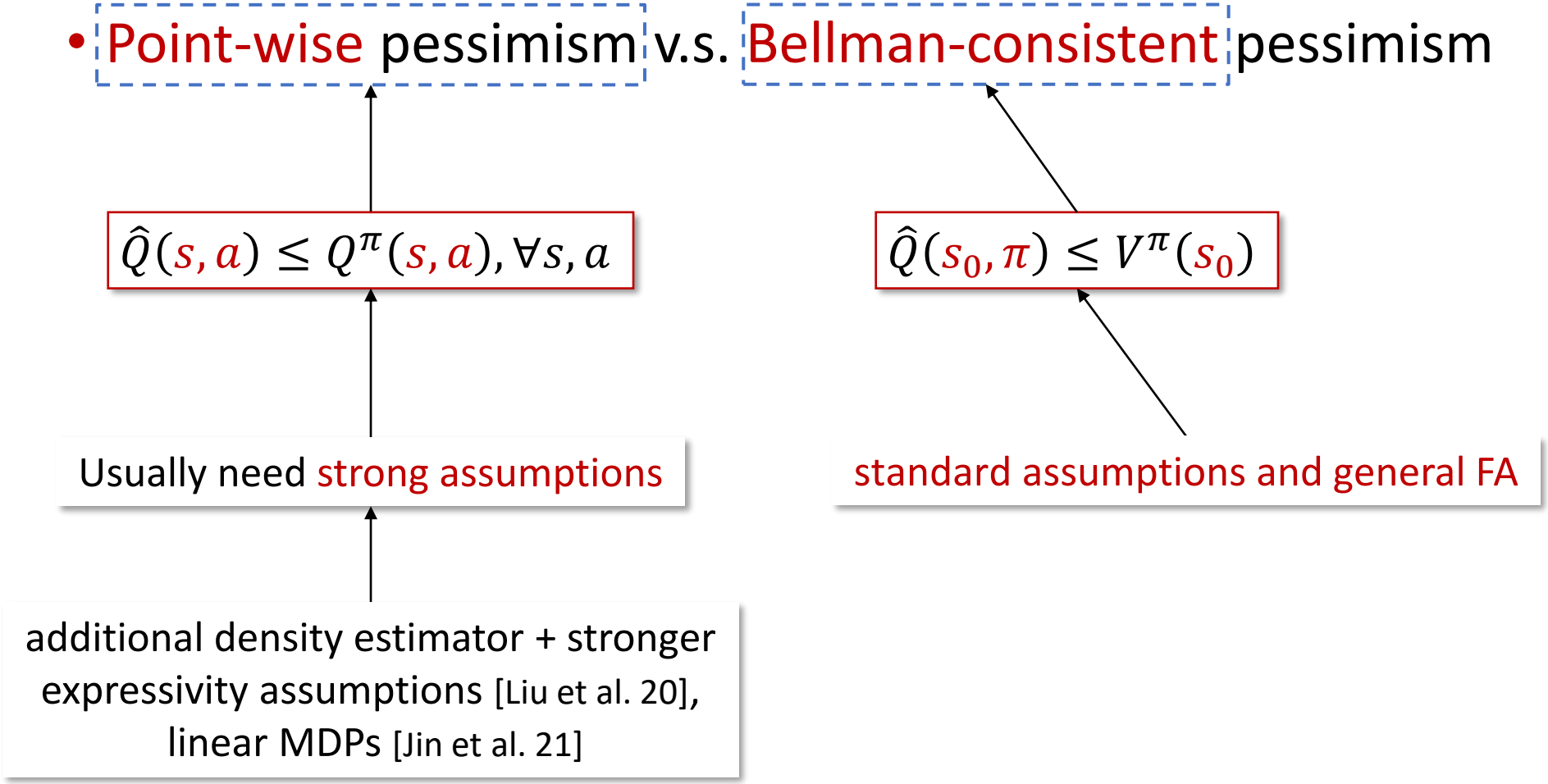
- Point-wise pessimism v.s. Bellman-consistent pessimism



[Liu et al 20] **Provably Good Batch Reinforcement Learning Without Great Exploration** NeurIPS'20
[Jin et al 21] **Is pessimism provably efficient for offline RL?** ICML'21

Comparison between Pessimism Frameworks

- Point-wise pessimism v.s. Bellman-consistent pessimism



[Liu et al. 20] **Provably Good Batch Reinforcement Learning Without Great Exploration** NeurIPS'20
[Jin et al. 21] **Is pessimism provably efficient for offline RL?** ICML'21

Results in Linear FA

MSPBE (quadratic form of θ) [Antos et al. 08]

- $f(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^T\theta,\ \theta \in \mathbb{R}^d$
- $\hat{\pi} = \underset{\pi\in\Pi}{\operatorname{argmax}} \min_{f\in\mathcal{F}_{\pi,\varepsilon}} f(s_0,\pi),\ \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F}: \mathbb{E}_{\mathcal{D}}[(f - \Pi\mathcal{T}^\pi f)^2] \leq \varepsilon$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma}\sqrt{\frac{d\log\left(\frac{dV_{\max}|\mathcal{A}|}{\delta}\right)}{n}}\,\mathbb{E}_{d_\pi}\left[\sqrt{\phi(s,a)^T\Sigma_{\mathcal{D}}^{-1}\phi(s,a)}\right]$$

[Antos et al. 08] **Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path**

Results in Linear FA

- $f(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^T\theta,\ \theta \in \mathbb{R}^d$
- $\hat{\pi} = \underset{\pi\in\Pi}{\operatorname{argmax}} \min_{f\in\mathcal{F}_{\pi,\varepsilon}} f(s_0,\pi),\ \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F}: \mathbb{E}_{\mathcal{D}}[(f - \Pi\mathcal{T}^\pi f)^2] \leq \varepsilon$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma}\sqrt{\frac{d\log\left(\frac{dV_{\max}|\mathcal{A}|}{\delta}\right)}{n}}\,\mathbb{E}_{d_\pi}\left[\sqrt{\phi(s,a)^T\Sigma_{\mathcal{D}}^{-1}\phi(s,a)}\right]$$

$\Sigma_{\mathcal{D}} := \mathbb{E}_{\mathcal{D}}[\phi(s,a)\phi(s,a)^T]$

Single-policy concentrability in linear FA

Results in Linear FA

- $f(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^T\theta,\ \theta \in \mathbb{R}^d$
- $\hat{\pi} = \underset{\pi\in\Pi}{\operatorname{argmax}} \min_{f\in\mathcal{F}_{\pi,\varepsilon}} f(s_0,\pi),\ \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F}: \mathbb{E}_{\mathcal{D}}[(f - \Pi\mathcal{T}^{\pi}f)^2] \leq \varepsilon$

Induced an additional $\log(|\mathcal{A}|)$ term compared with [Jin et al. 21] (resolved by [Zanette et al. 21])

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{V_{\max}}{1-\gamma}\sqrt{\frac{d\log\left(\frac{dV_{\max}|\mathcal{A}|}{\delta}\right)}{n}}\,\mathbb{E}_{d_\pi}\left[\sqrt{\phi(s,a)^T\Sigma_{\mathcal{D}}^{-1}\phi(s,a)}\right]$$

$d^2 \Longrightarrow d$ over bonus-based pessimism [Jin et al. 21]

[Jin et al. 21] **Is Pessimism Provably Efficient for Offline RL?** ICML 21
[Zanette et al. 2021] **Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning.** NeurIPS 21

Agenda

- Problem Setup
- Bellman-consistent Pessimism
- **Practical Algorithm**
- Related Works and Summary

Bottlenecks of Implementation

$$\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi), \qquad \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \leq \varepsilon\}$$

Issue 1: how to optimize policy over pessimistic PE?

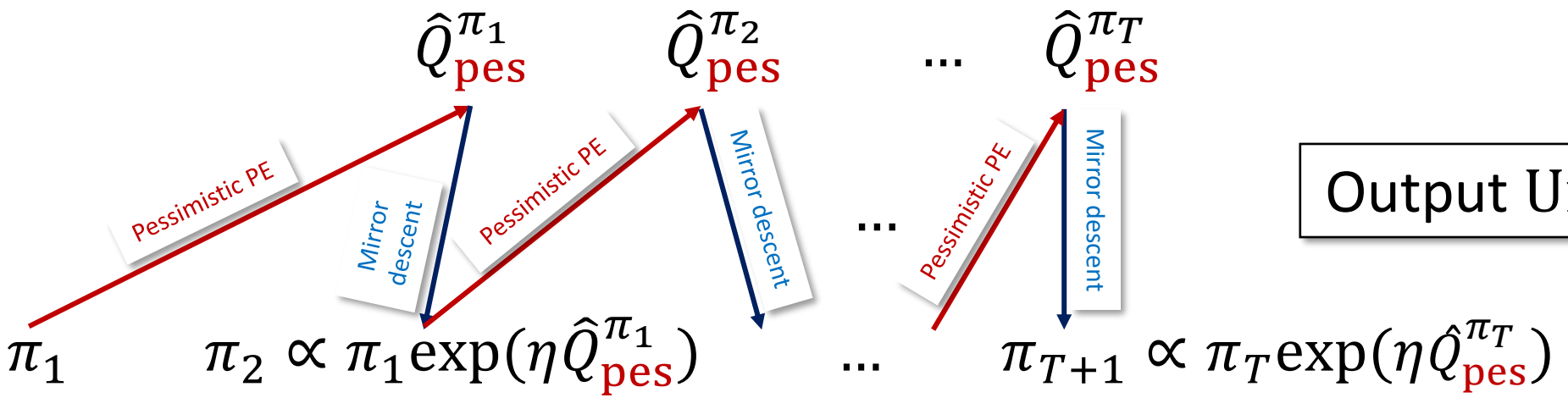
Bottlenecks of Implementation

$$\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmax}} \min_{f \in \mathcal{F}_{\pi,\varepsilon}} f(s_0, \pi), \qquad \mathcal{F}_{\pi,\varepsilon} \coloneqq \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \leq \varepsilon\}$$

Issue 2: how to find the most pessimistic evaluation over the version space?

1). Policy optimization---no-regret learning

- Mirror descent update



Output $\text{Unif}(\pi_{1:T})$

policy class $\Pi := \{\pi \propto \exp(\eta \sum_{i\in[t]} f^i), t \le T, f^{[t]} \in \mathcal{F}\}$

2). Pessimistic PE---constraint⟹regularization

- Constrained optimization

$$f_\pi \leftarrow \underset{f \in \mathcal{F}_{\pi,\varepsilon}}{\operatorname{argmin}} f(s_0, \pi), \qquad \mathcal{F}_{\pi,\varepsilon} := \{f \in \mathcal{F} : \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^\pi f)^2] \leq \varepsilon\}$$

- Regularized optimization

$$f_\pi \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} f(s_0, \pi) + \lambda \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^\pi f)^2]$$

Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For any π, any constant C, and any distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$

$$J(\pi) - J(\hat{\pi}) \lesssim \underbrace{\frac{C \cdot V_{\max}}{1-\gamma} \sqrt[3]{\frac{T \log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}}}_{\text{On-support error}} + \underbrace{\frac{V_{\max}}{1-\gamma} \sqrt{\frac{\log|\mathcal{A}|}{T}}}_{\text{Optimization error}} + \underbrace{\frac{1}{T} \sum_{t \in [T]} \frac{\langle d_\pi \setminus \nu, f_t - \mathcal{T}^\pi f_t \rangle}{1-\gamma}}_{\text{Off-support error}}$$

Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For any π, any constant C, and any distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$

Due to policy class complexity
(policy class $\Pi := \{\pi \propto \exp(\eta \sum_{i \in [t]} f^i), t \leq T, f^{[t]} \in \mathcal{F}\}$)

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{C \cdot V_{\max}}{1-\gamma} \sqrt[3]{\frac{T \log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}} + \frac{V_{\max}}{1-\gamma} \sqrt{\frac{\log|\mathcal{A}|}{T}} + \frac{1}{T} \sum_{t \in [T]} \frac{\langle d_\pi \setminus \nu, f_t - \mathcal{T}^\pi f_t \rangle}{1-\gamma}$$

Due to regularized PE

Theoretical Guarantee

- Performance guarantee (*simplified*, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$):

For any π, any constant C, and any distribution ν satisfies $\max_{f,f' \in \mathcal{F}} \frac{\|f - f'\|_{2,\nu}^2}{\|f - f'\|_{2,\mu}^2} \leq C$

$$J(\pi) - J(\hat{\pi}) \lesssim \frac{C \cdot V_{\max}}{1-\gamma} \sqrt[3]{\frac{T \log\left(\frac{|\mathcal{F}|}{\delta}\right)}{n}} + \frac{V_{\max}}{1-\gamma} \sqrt{\frac{\log|\mathcal{A}|}{T}} + \frac{1}{T} \sum_{t \in [T]} \frac{\langle d_\pi \setminus \nu, f_t - \mathcal{T}^\pi f_t \rangle}{1-\gamma}$$

$$:= \operatorname*{argmin}_{f \in \mathcal{F}} f(s_0, \pi_t) + \lambda \mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi_t} f)^2]$$

Full computational efficiency in linear FA

$$\mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \approx \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a) - r - \gamma f(s', \pi)\right)^2\right] - \min_{f' \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a) - r - \gamma f(s', \pi)\right)^2\right]$$

Full computational efficiency in linear FA

$$\mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \approx \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a) - r - \gamma f(s', \pi)\right)^2\right] - \min_{f' \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a) - r - \gamma f(s', \pi)\right)^2\right]$$

(In linear FA) $\Longrightarrow \mathbb{E}_{\mathcal{D}}\left[(f - \Pi \mathcal{T}^{\pi} f)^2\right] = \mathrm{MSPBE}$ [Proposition 2, Antos et al 08]

Projection operator

[Antos et al 08] **Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path**

Full computational efficiency in linear FA

$$\mathbb{E}_{\mathcal{D}}[(f - \mathcal{T}^{\pi} f)^2] \approx \mathbb{E}_{\mathcal{D}}\left[\left(f(s,a) - r - \gamma f(s',\pi)\right)^2\right] - \min_{f' \in \mathcal{F}} \mathbb{E}_{\mathcal{D}}\left[\left(f'(s,a) - r - \gamma f(s',\pi)\right)^2\right]$$

(In linear FA) $\Longrightarrow \mathbb{E}_{\mathcal{D}}\left[(f - \Pi\mathcal{T}^{\pi} f)^2\right] = \mathrm{MSPBE}$ [Proposition 2, Antos et al 08]

$f(s_0,\pi) + \lambda \|f - \Pi\mathcal{T}^{\pi} f\|^2_{2,\mathcal{D}} \longrightarrow$ Quadratic form $\quad (f(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^T\theta)$

Agenda

- Problem Setup
- Bellman-consistent Pessimism
- Practical Algorithm
- **Related Works and Summary**

Related Work

- Theoretical:
  - General FA with strong assumptions (e.g., density estimation) [Liu et al 2020]
  - Linear MDPs [Jin et al 2021], Linear FA w/ completeness [Zanette et al 2021]
  - Model-based [Uehara et al 2021]
  - …
- Empirical:
  - Model-free [Fujimoto et al 2019, Kumar et al 2020]
  - Model-based [Kidambi et al 2020, Yu et al 2020]
  - …

Conclusions

- Bellman-consistent Pessimism

1. A general framework for offline reinforcement learning
   - Applies to general function approximation
   - Simple and provably efficient algorithm with standard assumptions

2. Implementation-friendly practical algorithm
   - Fully computationally tractable in linear FA

Future Directions

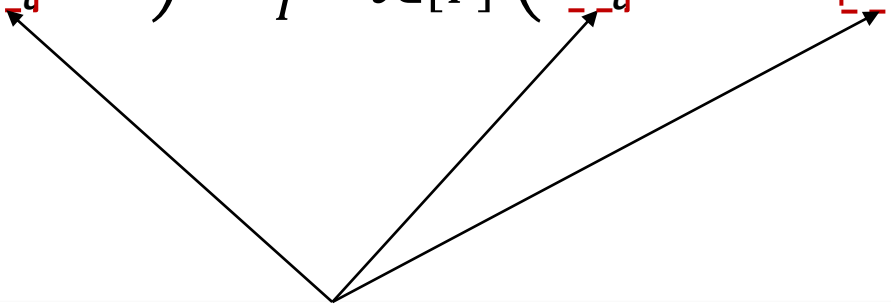
- Computationally tractable algorithm with $\tilde{O}(1/n^{0.5})$ rate
- Simply empirically practical algorithm
- Offline hyper-parameter tuning

Thank you!

Appendix

Proof Sketch of Practical Algorithm

- $J(\pi) - J(\hat{\pi}) \leq \frac{1}{T}\sum_{t\in[T]}\left(J(\pi) - J_{M_t}(\pi)\right) + \frac{1}{T}\sum_{t\in[T]}\left(J_{M_t}(\pi) - J_{M_t}(\pi_t)\right)$



An augmented MDP s.t.:

i) M_t has the same transition function as the true MDP

ii) f_t is the true Q^{π_t} in M_t

Proof Sketch of Practical Algorithm

- $J(\pi) - J(\hat{\pi}) \leq \frac{1}{T}\sum_{t\in[T]}\left(J(\pi) - J_{M_t}(\pi)\right) + \frac{1}{T}\sum_{t\in[T]}\left(J_{M_t}(\pi) - J_{M_t}(\pi_t)\right)$

$$= \frac{1}{T}\sum_{t\in[T]}\left(Q^{\pi}(s_0,\pi) - J_{M_t}(\pi)\right)$$

By performance difference Lemma $\longrightarrow$ $+\frac{1}{T}\sum_{t\in[T]}\mathbb{E}_{d_\pi}[f_t(s,\pi) - f_t(s,\pi_t)]$

Proof Sketch of Practical Algorithm

- $J(\pi) - J(\hat{\pi}) \leq \frac{1}{T}\sum_{t\in[T]}\left(J(\pi) - J_{M_t}(\pi)\right) + \frac{1}{T}\sum_{t\in[T]}\left(J_{M_t}(\pi) - J_{M_t}(\pi_t)\right)$

$$= \frac{1}{T}\sum_{t\in[T]}\left(Q^{\pi}(s_0,\pi) - J_{M_t}(\pi)\right)$$

$$+ \frac{1}{T}\sum_{t\in[T]}\mathbb{E}_{d_\pi}[f_t(s,\pi) - f_t(s,\pi_t)]$$

Controlled by $\mathbb{E}_{d_\pi}\left[\left(Q^{\pi} - \mathcal{T}^{\pi}_{M_t}Q^{\pi}\right)^2\right] = \mathbb{E}_{d_\pi}[(f_t - \mathcal{T}^{\pi_t}f_t)^2]$

Controlled by standard online learning argument