Pessimism in the Face of Confounders

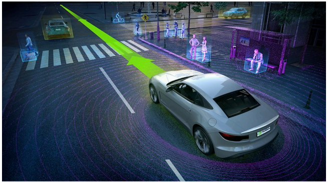
Provably Efficient Offline Reinforcement Learning in Partially Observable Markov Decision Processes

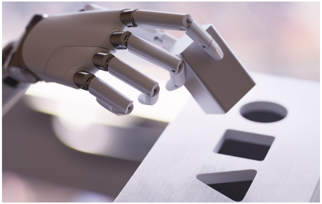
Miao Lu	Yifei Min	Zhaoran Wang	**Zhuoran Yang**
USTC	Yale U	Northwestern U	Yale U

Informs 2022

November 3, 2022

Offline RL: Learn to Plan from Offline Datasets

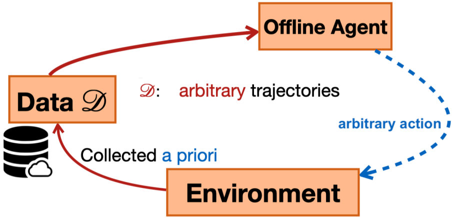




Offline RL: Learn how to plan from an **offline dataset** collected a priori, without any interaction with the environment.

images from internet sources.

Offline Policy Learning: Learn from Given Datasets

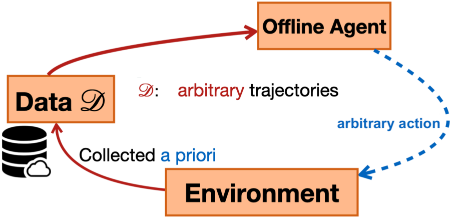


- Offline Data: collected a priori.
- Arbitrary trajectories: actions a_h by an offline agent (unknown rule).
- No further interactions with the environment
- Learning objective: performance of the learned policy

$$\mathrm{SubOpt}(\widehat{\pi}) = \sup_{\pi^\star \in \Pi} J(\pi^\star) - J(\widehat{\pi}),$$

where Π is a policy class, $\widehat{\pi}$ =OfflineRL$(\mathcal{D}, \mathcal{F})$.

Offline Policy Learning: Learn from Given Datasets



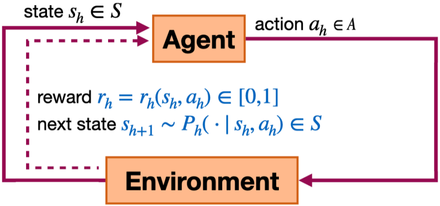
- Offline Data: collected a priori.
- Arbitrary trajectories: actions a_h by an offline agent (unknown rule).
- No further interactions with the environment
- Learning objective: performance of the learned policy

$$\mathrm{SubOpt}(\widehat{\pi}) = \sup_{\pi^\star \in \Pi} J(\pi^\star) - J(\hat{\pi}),$$

where Π is a policy class, $\widehat{\pi}$ =OfflineRL$(\mathcal{D}, \mathcal{F})$.

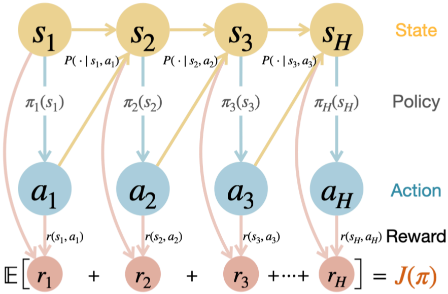
Offline RL in Partially Observable Markov Decision Processes

Episodic Markov Decision Process



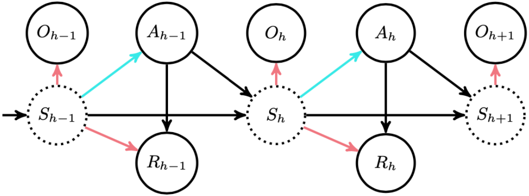
- $\mathcal{S}$: infinite state space. $\mathcal{A}$: finite action space.
- Unknown reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$.
- Unknown transition kernel $\mathbb{P}_h(\cdot \mid x, a) \in \Delta(\mathcal{S})$.
- Finite horizon H: terminate when $h = H$.

Episodic MDP



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

POMDP: "States" are Latent in MDP



- Latent state: $\{s_h\}_{h\in[H]}$ is unobserved.
- We observe an observation $o_h \sim \mathbb{O}_h(o \mid s_h) \in \Delta(\mathcal{O})$ emitted from latent state s_h.
- Observations: trajectory $\{(o_h, a_h, r_h), h \in [H]\}$.
- Reduced to Hidden Markov Model when $\{a_h\}_{h\in[H]}$ is fixed.

> Partial observability breaks Markov property
> $\Longrightarrow$ Consider history-dependent policy classes

History-Dependent Policy Class

- Observation after h-th step (partial trajectory): $\{(o_1, a_1), \cdots, (o_h, a_h)\}$
- History structure $\mathcal{H} = \{\mathcal{H}_h\}_{h=0}^{H-1}$:
  - each element $\tau_h \in \mathcal{H}_h$ is a (partial) trajectory
  - $\tau_h \subseteq \{(o_1, a_1), \cdots, (o_h, a_h)\}$
  - $\mathcal{H}_{h-1}$ reflects how much history we can look back into when determining a_h
  - $\mathcal{H}_{h-1}$ reflects memory constraint, chosen by algorithm, fixed
- $\mathcal{H}$-dependent policy $\Pi(\mathcal{H})$:

$$\pi \in \Pi(\mathcal{H}): \quad \pi_h(\cdot|o, \tau) : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A}), \forall h \in [H]$$

Goal: for a given $\mathcal{H}$, find the optimal $\pi^\star \in \Pi(\mathcal{H})$

$$\pi^\star \in \arg\max_{\pi \in \Pi(\mathcal{H})} J(\pi) := \arg\max_{\pi \in \Pi(\mathcal{H})} \mathbb{E}_\pi \left[\sum_{h=1}^{H} \gamma^{h-1} R_h(S_h, A_h) \right]$$

History-Dependent Policy Class

- Observation after h-th step (partial trajectory): $\{(o_1, a_1), \cdots, (o_h, a_h)\}$
- History structure $\mathcal{H} = \{\mathcal{H}_h\}_{h=0}^{H-1}$:
  - each element $\tau_h \in \mathcal{H}_h$ is a (partial) trajectory
  - $\tau_h \subseteq \{(o_1, a_1), \cdots, (o_h, a_h)\}$
  - $\mathcal{H}_{h-1}$ reflects how much history we can look back into when determining a_h
  - $\mathcal{H}_{h-1}$ reflects memory constraint, chosen by algorithm, fixed
- $\mathcal{H}$-dependent policy $\Pi(\mathcal{H})$:

$$\boxed{\pi \in \Pi(\mathcal{H}): \quad \pi_h(\cdot|o, \tau) : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A}),\ \forall h \in [H]}$$

Goal: for a given $\mathcal{H}$, find the optimal $\pi^\star \in \Pi(\mathcal{H})$

$$\pi^\star \in \arg\max_{\pi \in \Pi(\mathcal{H})} J(\pi) := \arg\max_{\pi \in \Pi(\mathcal{H})} \mathbb{E}_\pi \left[\sum_{h=1}^{H} \gamma^{h-1} R_h(S_h, A_h) \right]$$

History-Dependent Policy Class

- Observation after h-th step (partial trajectory): $\{(o_1, a_1), \cdots, (o_h, a_h)\}$
- History structure $\mathcal{H} = \{\mathcal{H}_h\}_{h=0}^{H-1}$:
  - each element $\tau_h \in \mathcal{H}_h$ is a (partial) trajectory
  - $\tau_h \subseteq \{(o_1, a_1), \cdots, (o_h, a_h)\}$
  - $\mathcal{H}_{h-1}$ reflects how much history we can look back into when determining a_h
  - $\mathcal{H}_{h-1}$ reflects memory constraint, chosen by algorithm, fixed
- $\mathcal{H}$-dependent policy $\Pi(\mathcal{H})$:

$$\pi \in \Pi(\mathcal{H}): \quad \pi_h(\cdot|o, \tau) : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A}), \forall h \in [H]$$

Goal: for a given $\mathcal{H}$, find the optimal $\pi^\star \in \Pi(\mathcal{H})$

$$\pi^\star \in \arg\max_{\pi \in \Pi(\mathcal{H})} J(\pi) := \arg\max_{\pi \in \Pi(\mathcal{H})} \mathbb{E}_\pi \left[\sum_{h=1}^{H} \gamma^{h-1} R_h(S_h, A_h) \right]$$

Examples of $\Pi(\mathcal{H})$

In our work, we consider three kinds of $\mathcal{H}$

- Reactive policy [Azizzadenesheli et al., 2018]: $\mathcal{H}_h = \{\emptyset\}$
- Finite-history policy [Efroni et al., 2022]: $\mathcal{H}_h = (\mathcal{O} \times \mathcal{A})^{\otimes \min\{k,h\}}$
- Hull-history policy [Liu et al., 2022]: $\mathcal{H}_h = (\mathcal{O} \times \mathcal{A})^{\otimes h}$

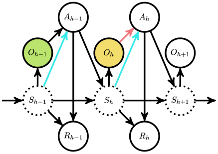


Figure: Reactive policy

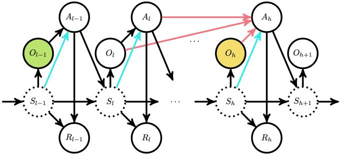


Figure: Finite-history policy

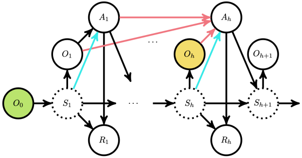
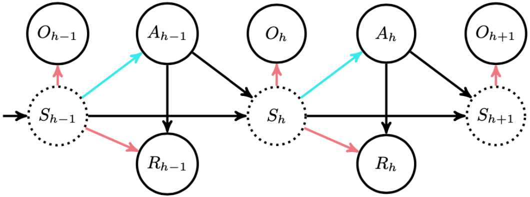


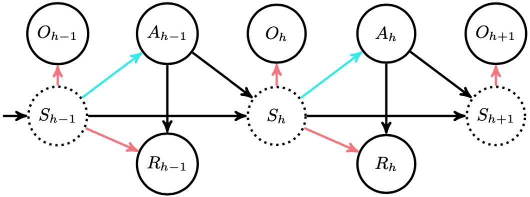
Figure: Full-history policy

Recap: Markov Policy and History-Dependent Policy



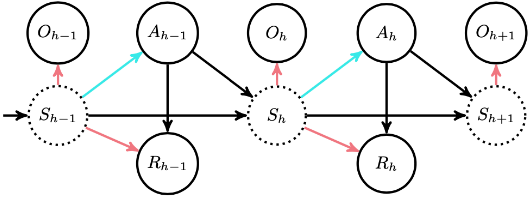
- Markov policy: $\pi = \{\pi_h\}_{h\in[H]}$
  - $\pi_h(\cdot|s)\colon \mathcal{S} \to \Delta(\mathcal{A})$
  - Have access to the latent state s_h (**unobserved**)
  - π induces a Markov chain $\{(s_h, a_h, o_h)\}_{h\in[H]}$
- History-dependent policy: $\pi = \{\pi_h\}_{h\in[H]}$
  - $\pi_h(\cdot|o,\tau) : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A})$, $\forall h \in [H]$
  - $\mathcal{H}_{h-1}$ reflects the memory size and complexity of π
  - π only involve observable quantities
- Goal: learn the optimal history-dependent policy within $\Pi(\mathcal{H})$.

Recap: Markov Policy and History-Dependent Policy



- Markov policy: $\pi = \{\pi_h\}_{h\in[H]}$
  - $\pi_h(\cdot|s) \colon \mathcal{S} \to \Delta(\mathcal{A})$
  - Have access to the latent state s_h (**unobserved**)
  - π induces a Markov chain $\{(s_h, a_h, o_h)\}_{h\in[H]}$
- History-dependent policy: $\pi = \{\pi_h\}_{h\in[H]}$
  - $\pi_h(\cdot|o,\tau) : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A})$, $\forall h \in [H]$
  - $\mathcal{H}_{h-1}$ reflects the memory size and complexity of π
  - π only involve observable quantities
- Goal: learn the optimal history-dependent policy within $\Pi(\mathcal{H})$.

Offline RL in POMDP – Data Generation



- Offline dataset $\mathbb{D}$ generated by a Markovian behavior policy π^b
- Observe n trajectories from π^b ($n =$ sample size)

$$\mathbb{D} = \left\{(o_0^k, (o_1^k, a_1^k, r_1^k), \cdots, (o_H^k, a_H^k, r_H^k))\right\}_{k=1}^n$$

- π^b generates (s_h, a_h, o_h, r_h) at each step, but s_h is not recorded in the dataset
- Motivation:
  - Healthcare records: s: full information; a: prescription; o: record
  - Autodriving: s: visual input; a: steering wheel; o: sensor data

Three Coupled Challenges:

- Confounding Issue (unique to POMDP)
- Insufficient Coverage
- Rich Observations

Three Coupled Challenges

- Confounding bias (o and r both depend on latent state)
  - $o_h \sim \mathbb{O}_h(\cdot \mid s_h)$
  - $r_h = r_h(s_h, a_h)$
  - s_h confounds r and o
  - pretending o_h is state and running standard RL methods lead to bias
- Insufficient coverage ($\mathcal{P}^{\pi^b}$ and $\mathcal{P}^{\pi}$ ($\pi \in \Pi(\mathcal{H})$) doesn't match)
  - also known as distributional shift
  - between trajectory of π^b and a family of trajectories
  - appear in offline RL w/o partial obs.
  - existing works has proposed methods based on pessimism principle.
- Large observation spaces (space $\mathcal{O}$ can be large or even infinite)
  - Need to incorporate function approximation tools.

Three Coupled Challenges

- Confounding bias (o and r both depend on latent state)
  - $o_h \sim \mathbb{O}_h(\cdot \mid s_h)$
  - $r_h = r_h(s_h, a_h)$
  - s_h confounds r and o
  - pretending o_h is state and running standard RL methods lead to bias
- Insufficient coverage ($\mathcal{P}^{\pi^b}$ and $\mathcal{P}^{\pi}$ ($\pi \in \Pi(\mathcal{H})$ doesn't match)
  - also known as distributional shift
  - between trajectory of π^b and a family of trajectories
  - appear in offline RL w/o partial obs.
  - existing works has proposed methods based on pessimism principle.
- Large observation spaces (space $\mathcal{O}$ can be large or even infinite)
  - Need to incorporate function approximation tools.

Three Coupled Challenges

- Confounding bias (o and r both depend on latent state)
  - $o_h \sim \mathbb{O}_h(\cdot \,|\, s_h)$
  - $r_h = r_h(s_h, a_h)$
  - s_h confounds r and o
  - pretending o_h is state and running standard RL methods lead to bias
- Insufficient coverage ($\mathcal{P}^{\pi^b}$ and $\mathcal{P}^{\pi}$ ($\pi \in \Pi(\mathcal{H})$ doesn't match)
  - also known as distributional shift
  - between trajectory of π^b and a family of trajectories
  - appear in offline RL w/o partial obs.
  - existing works has proposed methods based on pessimism principle.
- Large observation spaces (space $\mathcal{O}$ can be large or even infinite)
  - Need to incorporate function approximation tools.

Existing Works

Our work addresses all these challenges **simultaneously**.

- Offline RL in MDP: No partial obs. and no confouding issue.
- Online RL in POMDP: no distributional shift and confounding
- Offline Policy evaluation: easier distributional shift (π^b and π^e) and no policy learning

	Offline	Partial Obs.	Confound	Policy Opt.
Xie et al. [2021]	✓	✘	✘	✓
Uehara and Sun [2021]	✓	✘	✘	✓
Jin et al. [2020]	✘	✓	✘	✓
Liu et al. [2022]	✘	✓	✘	✓
Bennett and Kallus [2021]	✓	✓	✓	✘
Shi et al. [2021]	✓	✓	✓	✘
P3O (ours)	✓	✓	✓	✓

Our Algorithm: Proxy variable Pessimistic Policy Optimization (P3O)

Proxy variable Pessimistic Policy Optimization (P3O)

- Pessimistic policy optimization $\widehat{\pi} := \arg\max_{\pi \in \Pi(\mathcal{H})} \widehat{J}_{\mathsf{Pess}}(\pi)$
  - pessimism principle: $\widehat{J}_{\mathsf{Pess}}(\pi) \leq J(\pi)$ for all $\pi \in \Pi(\mathcal{H})$
- Policy value identification via proximal causal inference (PCI)
  - $J(\pi)$ is identified via value bridge functions $\mathbf{b}^\pi = \{b_h^\pi\}_{h \in [H]}$
  - $b_h^\pi : \mathcal{H}_{h-1} \times \mathcal{O} \to [0, H]$
  - $J(\pi) = \sum_a b_1^\pi(o_1, a)$
  - $\{b_h^\pi\}_{h \in [H]}$ satisfy Bellman-type moment equations
- Minimax estimation with uncertainty quantification
  - b_h^π's moment equation leads to a minimax estimation loss function
  - Construct a high-prob. confidence region $\mathrm{CR}^\pi(\xi)$ for $\mathbf{b}^\pi$
  - $\mathrm{CR}^\pi(\xi)$ constructed via sublevel sets of loss function
  - $\widehat{J}_{\mathsf{Pess}}(\pi) = \inf_{\mathbf{b} \in \mathrm{CR}^\pi(\xi)} \sum_a b_1(o_1, a)$

Proxy variable Pessimistic Policy Optimization (P3O)

- Pessimistic policy optimization $\widehat{\pi} := \arg\max_{\pi \in \Pi(\mathcal{H})} \widehat{J}_{\mathsf{Pess}}(\pi)$
  - pessimism principle: $\widehat{J}_{\mathsf{Pess}}(\pi) \leq J(\pi)$ for all $\pi \in \Pi(\mathcal{H})$
- Policy value identification via proximal causal inference (PCI)
  - $J(\pi)$ is identified via value bridge functions $\mathbf{b}^\pi = \{b_h^\pi\}_{h \in [H]}$
  - $b_h^\pi : \mathcal{H}_{h-1} \times \mathcal{O} \to [0, H]$
  - $J(\pi) = \sum_a b_1^\pi(o_1, a)$
  - $\{b_h^\pi\}_{h \in [H]}$ satisfy Bellman-type moment equations
- Minimax estimation with uncertainty quantification
  - b_h^π's moment equation leads to a minimax estimation loss function
  - Construct a high-prob. confidence region $\mathrm{CR}^\pi(\xi)$ for $\mathbf{b}^\pi$
  - $\mathrm{CR}^\pi(\xi)$ constructed via sublevel sets of loss function
  - $\widehat{J}_{\mathsf{Pess}}(\pi) = \inf_{\mathbf{b} \in \mathrm{CR}^\pi(\xi)} \sum_a b_1(o_1, a)$

Proxy variable Pessimistic Policy Optimization (P3O)

- Pessimistic policy optimization $\widehat{\pi} := \arg\max_{\pi\in\Pi(\mathcal{H})} \widehat{J}_{\mathsf{Pess}}(\pi)$
  - pessimism principle: $\widehat{J}_{\mathsf{Pess}}(\pi) \le J(\pi)$ for all $\pi \in \Pi(\mathcal{H})$

- Policy value identification via proximal causal inference (PCI)
  - $J(\pi)$ is identified via value bridge functions $\mathbf{b}^\pi = \{b_h^\pi\}_{h\in[H]}$
  - $b_h^\pi : \mathcal{H}_{h-1} \times \mathcal{O} \to [0, H]$
  - $J(\pi) = \sum_a b_1^\pi(o_1, a)$
  - $\{b_h^\pi\}_{h\in[H]}$ satisfy Bellman-type moment equations

- Minimax estimation with uncertainty quantification
  - b_h^π's moment equation leads to a minimax estimation loss function
  - Construct a high-prob. confidence region $\mathrm{CR}^\pi(\xi)$ for $\mathbf{b}^\pi$
  - $\mathrm{CR}^\pi(\xi)$ constructed via sublevel sets of loss function
  - $\widehat{J}_{\mathsf{Pess}}(\pi) = \inf_{\mathbf{b}\in\mathrm{CR}^\pi(\xi)} \sum_a b_1(o_1, a)$

P3O: Pessimistic Policy Optimization

P3O outputs the policy that maximizes the pessimistic estimator of $J(\pi)$:

$$\widehat{\pi} := \mathop{\arg\max}_{\pi \in \Pi(\mathcal{H})} \widehat{J}_{\mathsf{Pess}}(\pi), \quad \text{where} \quad \widehat{J}_{\mathsf{Pess}}(\pi) = \min_{\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)} \Big\{ \sum_{a \in \mathcal{A}} b_1(o_1, a) \Big\}$$

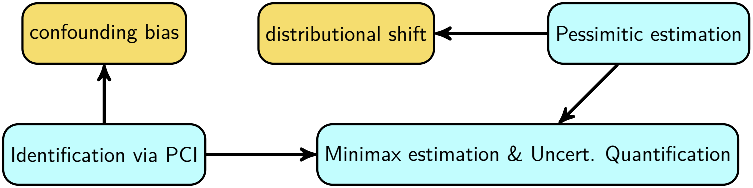
- (Policy evaluation) Construct confidence region $\mathrm{CR}^{\pi}(\xi)$ for the value bridge function $\{b_h^{\pi}\}_{h=1}^{H}$ via minimax estimation.
- (Policy optimization) Choose $\hat{\pi}$ that maximizes the pessimistic value estimator $\hat{J}_{\mathsf{Pess}}(\pi)$.

P3O: Pessimistic Policy Optimization

P3O outputs the policy that maximizes the pessimistic estimator of $J(\pi)$:

$$\widehat{\pi} := \arg\max_{\pi \in \Pi(\mathcal{H})} \widehat{J}_{\mathsf{Pess}}(\pi), \quad \text{where} \quad \widehat{J}_{\mathsf{Pess}}(\pi) = \min_{\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)} \Big\{ \sum_{a \in \mathcal{A}} b_1(o_1, a) \Big\}$$

- (Policy evaluation) Construct confidence region $\mathrm{CR}^{\pi}(\xi)$ for the value bridge function $\{b_h^{\pi}\}_{h=1}^{H}$ via minimax estimation.
- (Policy optimization) Choose $\hat{\pi}$ that maximizes the pessimistic value estimator $\hat{J}_{\mathsf{Pess}}(\pi)$.



Thank You!

Paper: Pessimism in the Face of Confounders: Provably Efficient Offline Reinforcement Learning in Partially Observable Markov Decision Processes `arxiv:2205.13589`

References I

K. Azizzadenesheli, Y. Yue, and A. Anandkumar. Policy gradient in partially observable environments: Approximation and convergence. arXiv preprint arXiv:1810.07900, 2018.

A. Bennett and N. Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. arXiv preprint arXiv:2110.15332, 2021.

Y. Efroni, C. Jin, A. Krishnamurthy, and S. Miryoosefi. Provable reinforcement learning with a short-term memory. arXiv preprint arXiv:2202.03983, 2022.

C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete pomdps. Advances in Neural Information Processing Systems, 33:18530–18539, 2020.

References II

Q. Liu, A. Chung, C. Szepesvári, and C. Jin. When is partially observable reinforcement learning not scary? arXiv preprint arXiv:2204.08967, 2022.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika, 105(4):987–993, 2018.

C. Shi, M. Uehara, and N. Jiang. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. arXiv preprint arXiv:2111.06784, 2021.

M. Uehara and W. Sun. Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. arXiv e-prints, pages arXiv–2107, 2021.

References III

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal.
Bellman-consistent pessimism for offline reinforcement learning.
Advances in neural information processing systems, 34, 2021.