Towards Optimal Offline Reinforcement Learning

Mengmeng Li[1] & Daniel Kuhn[1] & **Tobias Sutter**[2]

Fifth Symposium on Machine Learning and Dynamical Systems
Kyoto, February 9, 2026

[1] Risk Analytics and Optimization Chair, EPFL
[2]University of St.Gallen

EPFL

Swiss Institute for
Empirical Economic Research
University of St.Gallen

arXiv:2503.12283

Markov Decision Process

MDP $(\mathcal{S}, \mathcal{A}, r, P)$

- Finite spaces $\mathcal{S}, \mathcal{A}$
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- $S_{t+1} \sim P(\cdot|S_t, A_t)$, $S_0 = s$
  P unknown
- $\pi(\cdot|s') = \text{Prob}(A_t \in \cdot|S_t = s')$

Markov Decision Process

MDP $(\mathcal{S}, \mathcal{A}, r, P)$

- Finite spaces $\mathcal{S}, \mathcal{A}$
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- $S_{t+1} \sim P(\cdot|S_t, A_t)$, $S_0 = s$
  P unknown
- $\pi(\cdot|s') = \text{Prob}(A_t \in \cdot | S_t = s')$

Value function

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

Optimal policy

$$\pi^\star = \arg\max_{\pi} V_{\pi,P}$$

Markov Decision Process

MDP $(\mathcal{S}, \mathcal{A}, r, P)$

- Finite spaces $\mathcal{S}, \mathcal{A}$
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- $S_{t+1} \sim P(\cdot|S_t, A_t)$, $S_0 = s$
  P unknown
- $\pi(\cdot|s') = \text{Prob}(A_t \in \cdot | S_t = s')$

Value function

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

Optimal policy

$$\pi^\star = \arg\max_{\pi} V_{\pi,P}$$

Offline data: $\{S_t, A_t\}_{t=1}^{T}$ where $A_t \sim \pi_b(\cdot|S_t)$

- Behavioural policy π_b $\neq$ Evaluation policy π_e

Markov Decision Process

MDP $(\mathcal{S}, \mathcal{A}, r, P)$

- Finite spaces $\mathcal{S}, \mathcal{A}$
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- $S_{t+1} \sim P(\cdot|S_t, A_t)$, $S_0 = s$
  P unknown
- $\pi(\cdot|s') = \text{Prob}(A_t \in \cdot | S_t = s')$

Value function

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

Optimal policy

$$\pi^\star = \arg\max_{\pi} V_{\pi,P}$$

Offline data: $\{S_t, A_t\}_{t=1}^T$ where $A_t \sim \pi_b(\cdot|S_t)$

- Behavioural policy π_b $\neq$ Evaluation policy π_e
- **Off-policy evaluation problem (OPE):** Estimate $V_{\pi_e,P}$

Markov Decision Process

MDP $(\mathcal{S}, \mathcal{A}, r, P)$

- Finite spaces $\mathcal{S}, \mathcal{A}$
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- $S_{t+1} \sim P(\cdot|S_t, A_t)$, $S_0 = s$
  P unknown
- $\pi(\cdot|s') = \text{Prob}(A_t \in \cdot | S_t = s')$

Value function

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

Optimal policy

$$\pi^\star = \arg\max_{\pi} V_{\pi,P}$$

Offline data: $\{S_t, A_t\}_{t=1}^T$ where $A_t \sim \pi_b(\cdot|S_t)$

- Behavioural policy π_b $\neq$ Evaluation policy π_e
- **Off-policy evaluation problem (OPE):** Estimate $V_{\pi_e,P}$
- **Optimal policy learning:** Estimate $\pi^\star$

Off-policy evaluation

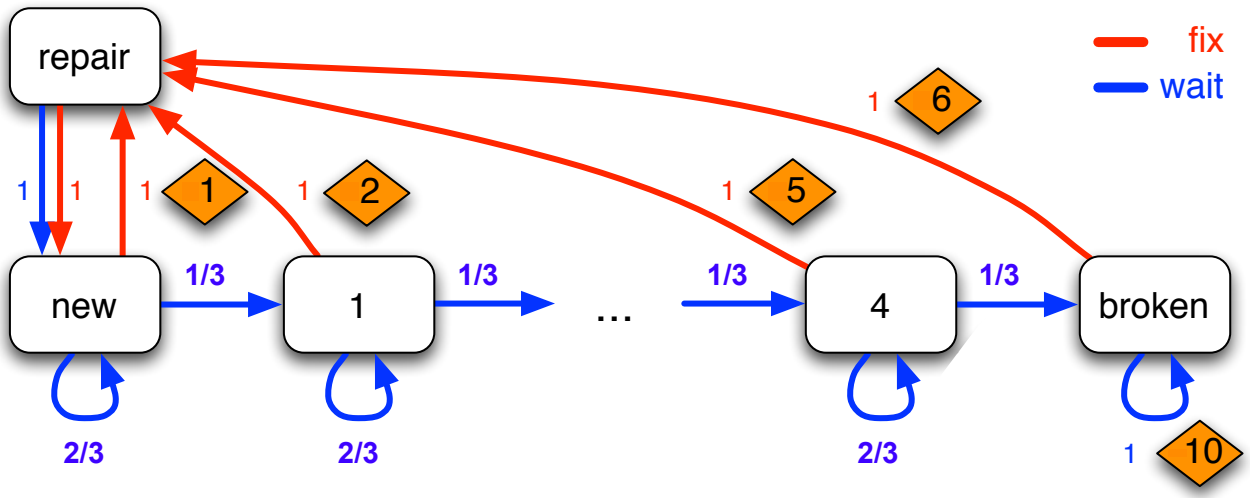
- **Value function**

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

- **Potential outcome framework**
  - [Athey and Wager, 2017, 2021], [Kitagawa and Tetenov, 2018], [Huber, 2021], [Chernozhukov et al., 2024], ...
- **Direct methods**
  - [Lagoudakis and Parr, 2003], [Ernst et al., 2005], [Mannor et al., 2007], [Antos et al., 2008], [Le et al., 2019], ...
- **Importance-weighting methods**
  - [Precup et al., 2000], [Swaminathan and Joachims, 2015], [Hirano et al., 2003], ...
- **Doubly-robust estimators**
  - [Jiang and Li, 2016], [Thomas and Brunskill, 2016], ...

OPE - Motivation - Machine Replacement

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$



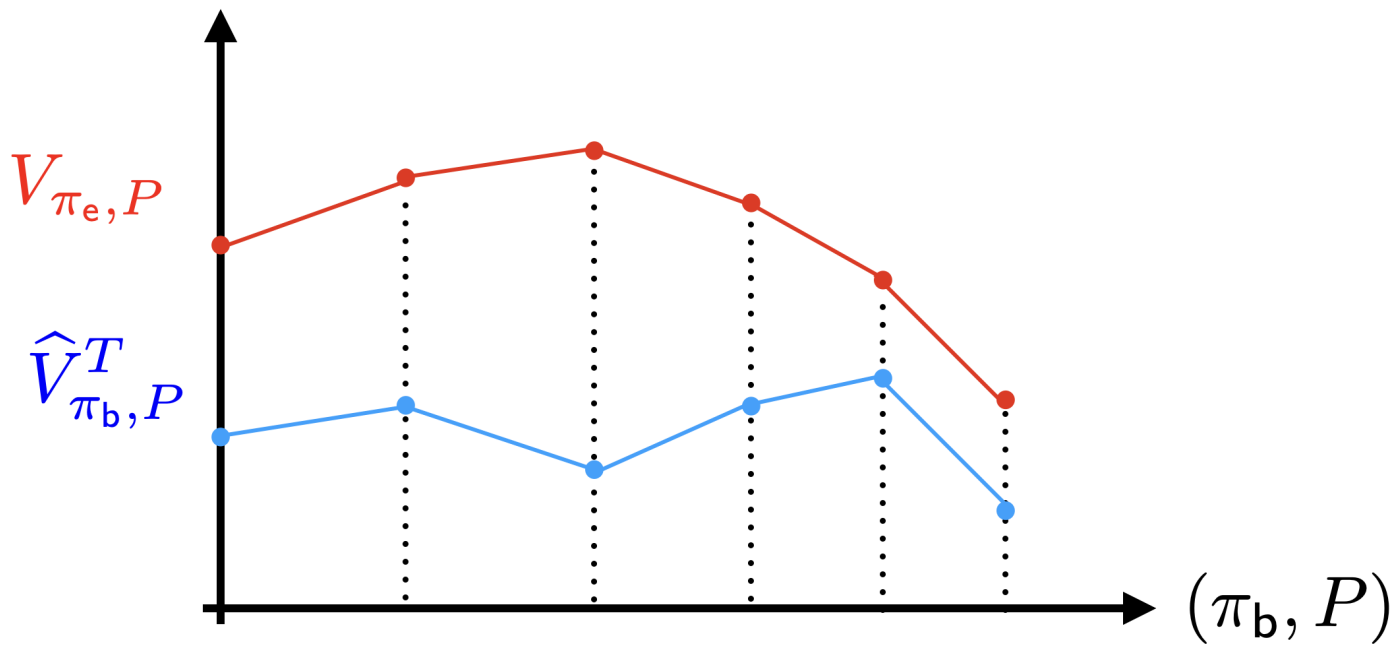
- $\mathcal{S} = \{\text{repair}, \text{new}, 1, 2, 3, 4, \text{broken}\}, \quad \mathcal{A} = \{\text{fix}, \text{wait}\}$
- Transition probability matrix P given by arrows
- Negative rewards $-r$ given by diamonds

OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^{T}$, $A_t \sim \pi_{\mathsf{b}}(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^{T}_{\pi_{\mathsf{b}},P}$ for $V_{\pi_{\mathsf{e}},P}$

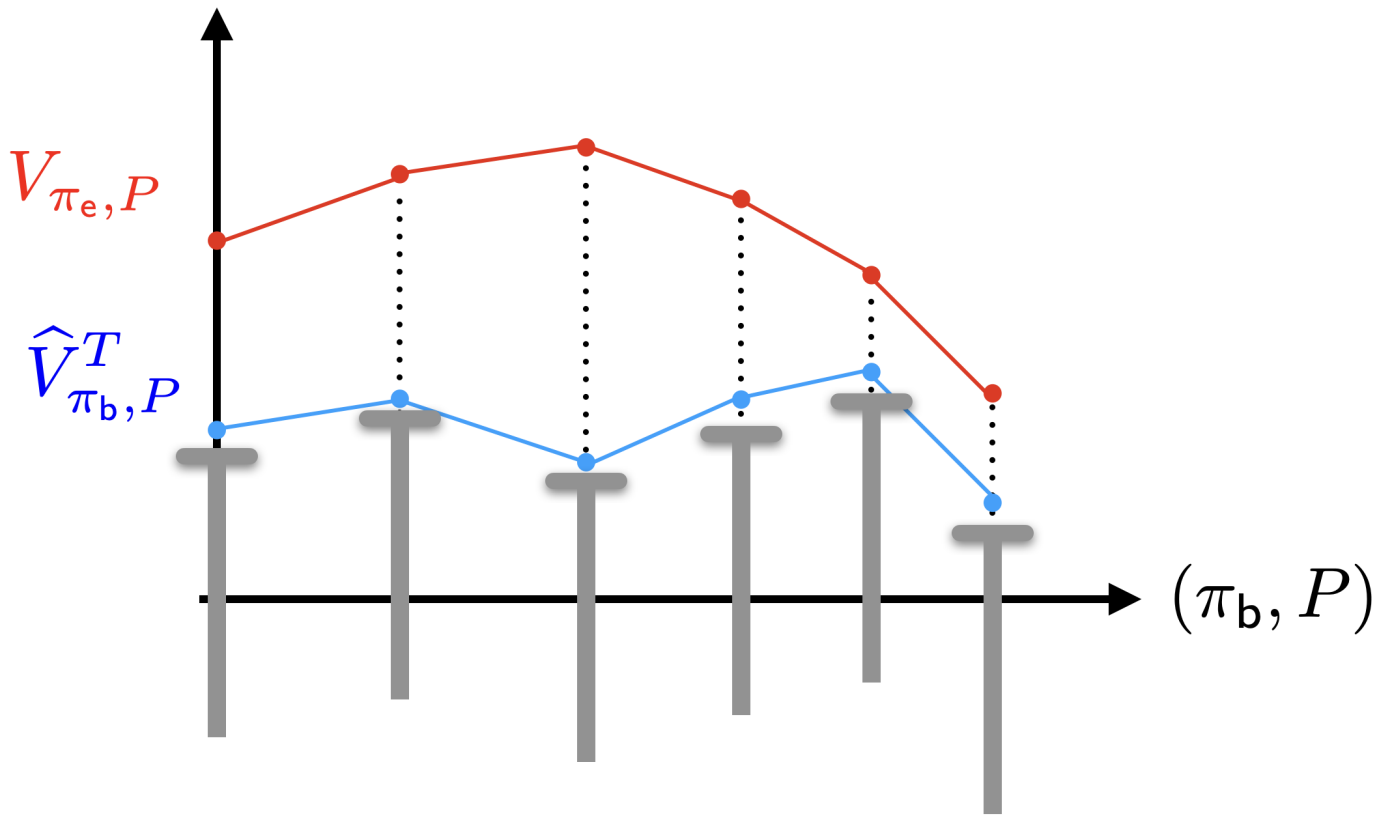
OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$



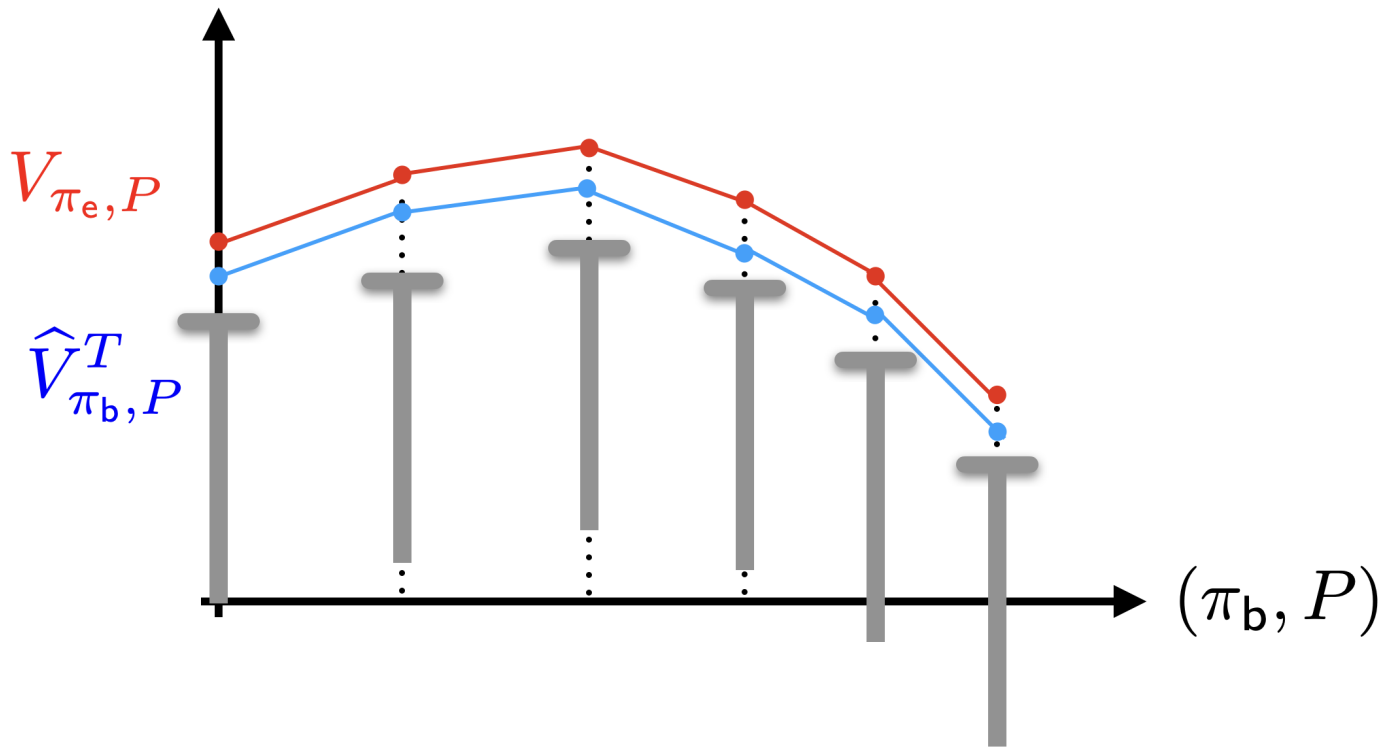
OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$



OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$



OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$

① High-probability under-estimator ($\rho > 0$):

$$\mathbb{P}_{\pi_b,P}\Big(\underbrace{V_{\pi_e,P} < \widehat{V}^T_{\pi_b,P}}_{\widehat{V}^T_{\pi_b,P} \text{ over-estimates } V_{\pi_e,P}} \Big) \leq e^{-\rho \cdot T} \quad \forall \pi_b, \forall P$$

OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$

(1) High-probability under-estimator ($\rho > 0$):

$$\mathbb{P}_{\pi_b,P}\Big(\underbrace{V_{\pi_e,P} < \widehat{V}^T_{\pi_b,P}}_{\widehat{V}^T_{\pi_b,P} \text{ over-estimates } V_{\pi_e,P}} \Big) \leq e^{-\rho \cdot T} \quad \forall \pi_b, \forall P$$

(2) Maximality/efficiency of estimator:
Any estimator $\overline{V}^T_{\pi_b,P}$ satisfying (1) for "large T" is such that

$$\widehat{V}^T_{\pi_b,P} \geq \overline{V}^T_{\pi_b,P} \quad \forall \pi_b, \forall P$$

OPE - Roadmap (cont'd)

- **Offline data:** $\{S_t, A_t\}_{t=1}^T$, $A_t \sim \pi_b(\cdot|S_t)$, $S_{t+1} \sim P(\cdot|S_t, A_t)$
- **Goal:** Build an **optimal estimator** $\widehat{V}^T_{\pi_b,P}$ for $V_{\pi_e,P}$

(1) High-probability under-estimator ($\rho > 0$):

$$\mathbb{P}_{\pi_b,P}\Big(\underbrace{V_{\pi_e,P} < \widehat{V}^T_{\pi_b,P}}_{\widehat{V}^T_{\pi_b,P} \text{ over-estimates } V_{\pi_e,P}} \Big) \leq e^{-\rho \cdot T} \quad \forall \pi_b, \forall P$$

(2) Maximality/efficiency of estimator:
Any estimator $\overline{V}^T_{\pi_b,P}$ satisfying (1) for "large T" is such that

$$\widehat{V}^T_{\pi_b,P} \geq \overline{V}^T_{\pi_b,P} \quad \forall \pi_b, \forall P$$

Difficulty: Behavioural policy π_b and transition kernel P are unknown

Reparametrization

- $\xi \in \Xi := \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ unique stationary state-action-next-state distribution induced by (π, P)

$$\xi\left(s, a, s'\right) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi, P}\left(S_t = s, A_t = a, S_{t+1} = s'\right)$$

Reparametrization

- $\xi \in \Xi := \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ unique stationary state-action-next-state distribution induced by (π, P)

$$\xi\left(s, a, s'\right) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi, P}\left(S_t = s, A_t = a, S_{t+1} = s'\right)$$

- Bijection $\xi \longleftrightarrow (\pi, P)$
  - $\xi_{\mathsf{b}} \longleftrightarrow (\pi_{\mathsf{b}}, P)$, $\xi_{\mathsf{e}} \longleftrightarrow (\pi_{\mathsf{e}}, P)$

Reparametrization

- $\xi \in \Xi := \Delta(\mathcal{S} \times \mathcal{A} \times \mathcal{S})$ unique stationary state-action-next-state distribution induced by (π, P)

$$\xi\left(s, a, s'\right) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi, P}\left(S_t = s, A_t = a, S_{t+1} = s'\right)$$

- Bijection $\xi \longleftrightarrow (\pi, P)$
  - $\xi_{\mathsf{b}} \longleftrightarrow (\pi_{\mathsf{b}}, P), \qquad \xi_{\mathsf{e}} \longleftrightarrow (\pi_{\mathsf{e}}, P)$
- Linear parametrization of the value function

$$\begin{aligned} V_{\pi, P} &= \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi, P}[r(S_t, A_t)] \\ &= \lim_{T \to \infty} \sum_{s, s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \frac{1}{T} \sum_{t=1}^{T} \mathbb{P}_{\pi, P}\left(S_t = s, A_t = a, S_{t+1} = s'\right) \\ &= \sum_{s, s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} r(s, a) \xi\left(s, a, s'\right) \\ &=: V(\xi) \end{aligned}$$

Distribution shift

Goal: Build an **optimal** estimator for $V_{\pi_e,P} = V(\xi_e)$

Distribution shift

Goal: Build an **optimal** estimator for $V_{\pi_e,P} = V(\xi_e)$

- For every (exploratory) evaluation policy π_e, we can define a **distribution shift function** f_{π_e} such that

$$f_{\pi_e}(\xi_b) = \xi_e$$

Distribution shift

Goal: Build an **optimal** estimator for $V_{\pi_e,P} = V(\xi_e)$

- For every (exploratory) evaluation policy π_e, we can define a **distribution shift function** f_{π_e} such that

$$f_{\pi_e}(\xi_b) = \xi_e$$

- f_{π_e} is well defined, since
  - P uniquely determined by ξ_b
  - ξ_e uniquely determined by (P, π_e)

Distribution shift

Goal: Build an **optimal** estimator for $V_{\pi_e,P} = V(\xi_e)$

- For every (exploratory) evaluation policy π_e, we can define a **distribution shift function** f_{π_e} such that

$$f_{\pi_e}(\xi_b) = \xi_e$$

- f_{π_e} is well defined, since
  - P uniquely determined by ξ_b
  - ξ_e uniquely determined by (P, π_e)
- A natural choice is $V(\xi_e) = V(f_{\pi_e}(\xi_b))$
- ξ_b not known since P is unknown

$$\widehat{\xi}_T(s,a,s') = \frac{1}{T}\left(\sum_{t=1}^{T-1} 1_{S_t=s,A_t=a,S_{t+1}=s'} + 1_{S_T=s,A_T=a,S_1=s'}\right)$$

Distribution shift

Goal: Build an **optimal** estimator for $V_{\pi_e,P} = V(\xi_e)$

- For every (exploratory) evaluation policy π_e, we can define a **distribution shift function** f_{π_e} such that

$$f_{\pi_e}(\xi_b) = \xi_e$$

- f_{π_e} is well defined, since
  - P uniquely determined by ξ_b
  - ξ_e uniquely determined by (P, π_e)
- A natural choice is $V(\xi_e) = V(f_{\pi_e}(\xi_b))$
- ξ_b not known since P is unknown

$$\widehat{\xi}_T\left(s,a,s'\right) = \frac{1}{T}\left(\sum_{t=1}^{T-1} 1_{S_t=s,A_t=a,S_{t+1}=s'} + 1_{S_T=s,A_T=a,S_1=s'}\right)$$

- Plug-in estimator (direct method): $V(f_{\pi_e}(\widehat{\xi}_T))$

Conditional relative entropy

Definition The conditional relative entropy of $\xi' \in \Xi$ with respect to $\xi \in \Xi$ is defined as

$$\mathrm{D}_{\mathrm{mdp}}\left(\xi' \| \xi\right) = \sum_{s,s' \in \mathcal{S}, a \in \mathcal{A}} \xi'\left(s,a,s'\right) \left(\log \frac{\xi'\left(s,a,s'\right)}{\sum_{\tilde{s} \in \mathcal{S}, \tilde{a} \in \mathcal{A}} \xi'(s,\tilde{a},\tilde{s})} - \log \frac{\xi\left(s,a,s'\right)}{\sum_{\tilde{s} \in \mathcal{S}, \tilde{a} \in \mathcal{A}} \xi(s,\tilde{a},\tilde{s})} \right)$$

- Used to quantify the "distance" between ξ' and ξ
- If $\xi' = \xi$, then $\mathrm{D}_{\mathrm{mdp}}(\xi' \| \xi) = 0$
- convex in ξ', non-convex in ξ

Optimal OPE estimator

$$V(\xi_e) = V_{\pi_e, P} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi_e, P}[r(S_t, A_t)]$$

Optimal OPE estimator

$$V(\xi_e) = V_{\pi_e,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi_e,P}[r(S_t, A_t)]$$

- Define a function

$$V_\rho{}^{\pi_e}\left(\xi'\right) = \inf_{\xi} \left\{ V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \leq \rho \right\}$$

Optimal OPE estimator

$$V(\xi_e) = V_{\pi_e, P} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi_e, P}[r(S_t, A_t)]$$

- Define a function

$$V_\rho{}^{\pi_e}\left(\xi'\right) = \inf_{\xi} \left\{ V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \leq \rho \right\}$$

- Distributionally robust estimator

$$\widehat{V}_{\xi_b}^{T} = V_\rho{}^{\pi_e}\left(\widehat{\xi}_T\right)$$

  - Data generated by $\mathbb{P}_{\xi_b}$

Optimal OPE estimator

$$V(\xi_e) = V_{\pi_e,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi_e,P}[r(S_t, A_t)]$$

- Define a function

$$V_\rho{}^{\pi_e}\left(\xi'\right) = \inf_{\xi} \left\{ V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \le \rho \right\}$$

- Distributionally robust estimator

$$\widehat{V}^T_{\xi_b} = V_\rho{}^{\pi_e}\left(\widehat{\xi}_T\right)$$

  - Data generated by $\mathbb{P}_{\xi_b}$

Theorem (High-probability under-estimator) For every ξ_b there is $\bar{c} > 0$ such that for every $\rho > 0$ and every π_e

$$\mathbb{P}_{\xi_b}\left(V(\xi_e) < \widehat{V}^T_{\xi_b}\right) \le e^{-\rho \cdot T} \cdot T e^{\bar{c}} (T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Optimal OPE estimator (cont'd)

DRO estimator

$$\widehat{V}_{\xi_b}^{T} = V_{\rho}^{\pi_e}\left(\widehat{\xi}_T\right)$$

① High-probability under-estimator:

$$\mathbb{P}_{\xi_b}\left(V(\xi_e) < \widehat{V}_{\xi_b}^{T}\right) \leq e^{-\rho \cdot T} \cdot Te^{\bar{c}}(T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Optimal OPE estimator (cont'd)

DRO estimator

$$\widehat{V}_{\xi_b}^{T} = V_{\rho}^{\pi_e}\left(\widehat{\xi}_T\right)$$

① High-probability under-estimator:

$$\mathbb{P}_{\xi_b}\left(V(\xi_e) < \widehat{V}_{\xi_b}^{T}\right) \leq e^{-\rho \cdot T} \cdot Te^{\bar{c}}(T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Theorem (Efficiency) For every π_e and every lsc function $\overline{V}^{\pi_e} : \Xi \to \mathbb{R}$ such that the estimator $\overline{V}^{\pi_e}(\widehat{\xi}_T)$ satisfies ①, we have

$$V_{\rho}^{\pi_e}\left(\xi'\right) \geq \overline{V}^{\pi_e}(\xi') \quad \forall \xi' \in \Xi$$

Optimal OPE estimator (cont'd)

DRO estimator

$$\widehat{V}_{\xi_b}^{T} = V_{\rho}^{\pi_e}\left(\widehat{\xi}_T\right)$$

① High-probability under-estimator:

$$\mathbb{P}_{\xi_b}\left(V(\xi_e) < \widehat{V}_{\xi_b}^{T}\right) \leq e^{-\rho \cdot T} \cdot Te^{\bar{c}}(T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Theorem (Efficiency) For every π_e and every lsc function $\overline{V}^{\pi_e} : \Xi \to \mathbb{R}$ such that the estimator $\overline{V}^{\pi_e}(\widehat{\xi}_T)$ satisfies ①, we have

$$V_{\rho}^{\pi_e}\left(\xi'\right) \geq \overline{V}^{\pi_e}(\xi') \quad \forall \xi' \in \Xi$$

Proposition (Consistency) Let $\{\rho_T\}$ be a non-negative sequence such that $\lim_{T\to\infty} \rho_T = 0$, then

$$\lim_{T\to\infty} V_{\rho_T}^{\pi_e}(\widehat{\xi}_T) = V(\xi_e) \quad \mathbb{P}_{\xi_b}\text{-a.s.} \quad \forall \xi_b$$

Computation of DRO estimator

- Optimal estimation given by function

$$V_{\rho}^{\pi_e}\left(\xi'\right) = \inf_{\xi}\left\{V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \leq \rho\right\}$$

  - Objective function only implicitly available
  - Feasible set non-convex

Computation of DRO estimator

- Optimal estimation given by function

$$V_\rho^{\pi_e}\left(\xi'\right) = \inf_{\xi}\left\{V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \leq \rho\right\}$$

  - Objective function only implicitly available
  - Feasible set non-convex

- Basic facts from our reparametrization
  - $\xi(s,a,s') = P(s' \mid s,a)\mu(s,a)$
  - $\mu(s,a) = \sum_{s' \in \mathcal{S}} \xi\left(s,a,s'\right)$
  - $P(s' \mid s,a) = \frac{\xi(s,a,s')}{\mu(s,a)}$

Computation of DRO estimator

- Optimal estimation given by function

$$V_\rho^{\pi_e}\left(\xi'\right) = \inf_{\xi}\left\{V\left(f_{\pi_e}\left(\xi\right)\right) : \mathsf{D}_{\mathsf{mdp}}\left(\xi' \| \xi\right) \leq \rho\right\}$$

  - Objective function only implicitly available
  - Feasible set non-convex

- Basic facts from our reparametrization
  - $\xi(s,a,s') = P(s' \mid s,a)\mu(s,a)$
  - $\mu(s,a) = \sum_{s' \in \mathcal{S}} \xi\left(s,a,s'\right)$
  - $P(s' \mid s,a) = \frac{\xi(s,a,s')}{\mu(s,a)}$

Lemma (Reformulation) Let $x = (s,a)$ and $\mathcal{X} = \mathcal{S} \times \mathcal{A}$,

$$V_\rho^{\pi_e}\left(\xi'\right) = \min_{P}\left\{\sum_{x \in \mathcal{X}} r(x)\mu_{\pi_e,P}(x) : \sum_{x \in \mathcal{X}} \mu'(x)\mathsf{D}(P'(\cdot \mid x) \| P(\cdot \mid x)) \leq \rho\right\}$$

where P' and μ' are induced by ξ'

Computation of DRO estimator (cont'd)

$$V_\rho^{\pi_e}\left(\xi'\right) = \min_{P}\left\{\sum_{x\in\mathcal{X}} r(x)\mu_{\pi_e,P}(x) : \sum_{x\in\mathcal{X}} \mu'(x)\mathrm{D}(P'(\cdot\,|\,x)\|P(\cdot\,|\,x)) \le \rho\right\}$$

- Non-convex objective, convex feasible set $\mathcal{P}_\rho(\xi')$

$$V_\rho^{\pi_e}\left(\xi'\right) = \min_{P\in\mathcal{P}_\rho(\xi')} \sum_{x\in\mathcal{X}} r(x)\mu_{\pi_e,P}(x)$$

- Interpretation as worst-case MDP
  - Literature on robust MDPs [Wiesemann et al., 2013], [Li et al., 2023], ...
  - Non-rectangularity of the uncertainty set

The critic

$$V_\rho^{\pi_e}\left(\xi'\right) = \min_{P \in \mathcal{P}_\rho(\xi')} \sum_{x \in \mathcal{X}} r(x)\mu_{\pi_e,P}(x) = \min_{\lambda \in \Lambda} V_{P^\lambda}^{\pi_e}$$

- Wlog can express $\mathcal{P}_\rho\left(\xi'\right) = \left\{P^\lambda : \lambda \in \Lambda\right\}$, where $\Lambda \subseteq \mathbb{R}^q$ is a solid, convex set, and P^λ is an affine function.

Algorithm: Projected Langevin dynamics
Require: Iterations $M \in \mathbb{N}$, step size $\eta > 0$, Gibbs parameter $\beta > 1$
Initialize $\lambda^{(0)} \in \Lambda$, $m = 0$
while $m \leq M - 1$ **do**
  Sample $w_{m+1} \sim \mathcal{N}\left(0, I_q\right)$
  Find $\lambda^{(m+1)} = \mathrm{Proj}_\Lambda\left(\lambda^{(m)} - \eta \nabla_\lambda V_{P^\lambda}^{\pi_e}|_{\lambda=\lambda^{(m)}} + \sqrt{2\eta/\beta} w_{m+1}\right)$
  $m \to m + 1$
end while

The critic (cont'd)

Theorem (Convergence - informal) If $\delta > 0$ $\exists$ universal constants $a > 4, b > 1$ and $c_1, c_2 > 0$ such that for all $\beta \geq c_1 q$ and $M \geq c_2 \cdot e^{q^b} \cdot \delta^{-a}$ the distribution ν_M of the output $\lambda^{(M)}$ of the algorithm satisfies

$$\mathbb{E}_{\lambda \sim \nu_M} \left[V_{P^\lambda}^{\pi_e} \right] \leq V_\rho^{\pi_e}(\xi') + \delta$$

- Exponential growth in $q = \dim \Lambda$ (curse of dimensionality) [Wiesemann et al., 2013]

Optimal policy learning

- Optimal policy

$$\pi^\star = \operatorname*{argmax}_{\pi} V_{\pi,P}$$

- Natural DRO estimator

$$\underbrace{\pi_\rho(\widehat{\xi}_T)}_{\widehat{\pi}_T} \in \operatorname*{argmax}_{\pi} V_\rho^\pi(\widehat{\xi}_T)$$

Theorem (High-probability under-estimator) For every ξ_b there is $\bar{c} > 0$ such that for every $\rho > 0$

$$\mathbb{P}_{\xi_b}\Big(\underbrace{V(f_{\pi_\rho(\widehat{\xi}_T)}(\xi_b))}_{V_{\widehat{\pi}_T,P}} < V_\rho^{\pi_\rho(\widehat{\xi}_T)}(\widehat{\xi}_T)\Big) \le e^{-\rho\cdot T}\cdot T e^{\bar{c}}(T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Optimal policy learning (cont'd)

DRO Estimators

$$\underbrace{\pi_\rho(\widehat{\xi}_T)}_{\widehat{\pi}_T} \in \operatorname*{argmax}_{\pi} V_\rho^{\pi}(\widehat{\xi}_T)$$

① High-probability under-estimator:

$$\mathbb{P}_{\xi_b}\Big(\underbrace{V(f_{\pi_\rho(\widehat{\xi}_T)}(\xi_b))}_{V_{\widehat{\pi}_T,P}} < V_\rho^{\pi_\rho(\widehat{\xi}_T)}(\widehat{\xi}_T)\Big) \le e^{-\rho\cdot T}\cdot Te^{\bar{c}}(T+1)^{(|\mathcal{S}||\mathcal{A}|)^2}$$

Efficiency (informal) For every pair $(\overline{\pi}(\cdot), \overline{V}^{\overline{\pi}(\cdot)}(\cdot))$ satisfying ①, we have

$$V_\rho^{\pi_\rho(\xi')}(\xi') \ge \overline{V}^{\overline{\pi}(\xi')}(\xi') \quad \forall \xi' \in \Xi$$

Computing an optimal policy

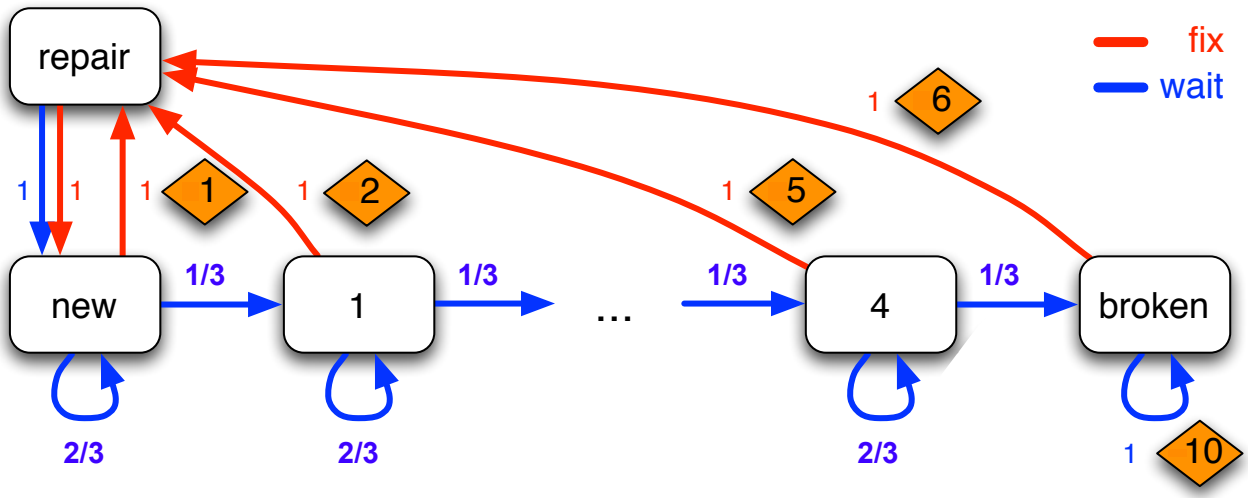
Algorithm: Actor-Critic
Require: Iteration number K, step size $\eta > 0$, tolerance $\delta > 0$
Initialize $\pi^{(0)}$, $k = 0$
while $k \leq K$ **do**
  Critic: Find $P^{(k)} \in \mathcal{P}_\rho(\xi')$ such that $V_{P^{(k)}}^{\pi^{(k)}} \leq V_\rho^{\pi^{(k)}}(\xi') + \delta$
  Actor: $\pi^{(k+1)} = \mathrm{Proj}_\Pi\left(\pi^{(k)} + \eta \nabla_\pi V_{P^{(k)}}^{\pi^{(k)}}\right)$
  $k \to k+1$
end while

Convergence rate:

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(\max_\pi V_\rho^\pi(\xi') - V_\rho^{\pi^{(k)}}(\xi')\right) = O(K^{-1/4})$$

Machine replacement problem

$$V_{\pi,P} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi,P}[r(S_t, A_t)]$$

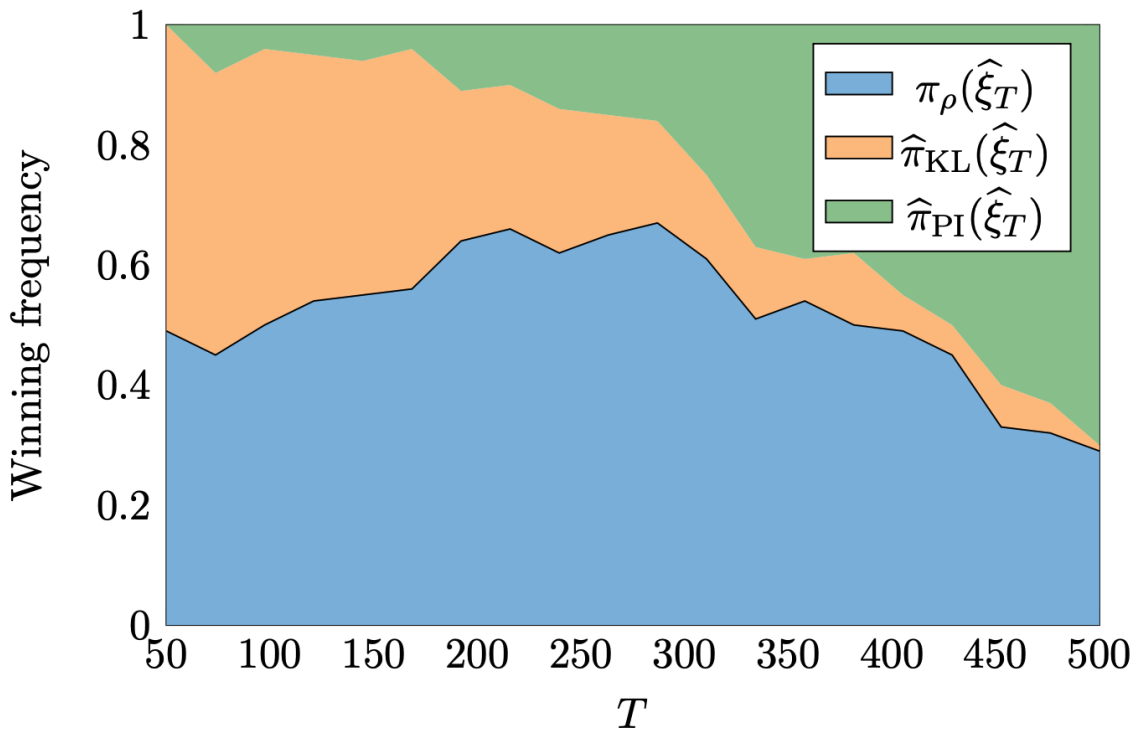


- $\mathcal{S} = \{\text{repair}, \text{new}, 1, 2, 3, 4, \text{broken}\}, \quad \mathcal{A} = \{\text{fix}, \text{wait}\}$
- Transition probability matrix P given by arrows
- Negative rewards $-r$ given by diamonds

Machine replacement problem - empirical results

Comparison of 3 different policies learnt from offline data

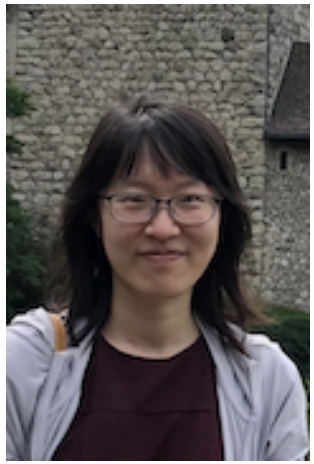
- $\pi_\rho(\widehat{\xi}_T)$
- KL-based approach $\widehat{\pi}_{\mathrm{KL}}(\widehat{\xi}_T)$ [Shi and Chi, 2024]
- Plug-in-based approach $\widehat{\pi}_{\mathrm{PI}}(\widehat{\xi}_T)$ [Li et al., 2024]



Summary

1. Offline RL approach for learning from a **single trajectory** of offline data with **unknown** behavioural policy π_b
2. Distributionally robust OPE estimator that is **statistically efficient**
3. Distributionally robust policy learning that is **statistically efficient**
4. Computation of estimators reduces to solving a **robust MDP** with **non-rectangular uncertainty set**
5. Numerical results show competitiveness of proposed approach

References



Mengmeng Li



Daniel Kuhn

This talk

- M. Li, D. Kuhn and T. Sutter, *Towards Optimal Offline Reinforcement Learning*, arXiv:2503.12283, 2025
- M. Li, D. Kuhn and T. Sutter, *Policy Gradient Algorithms for Robust MDPs with Non-Rectangular Uncertainty Sets*, SIAM Journal on Optimization, 2026