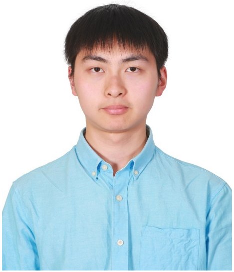
Minimax Off-Policy Evaluation for Multi-Armed Bandits

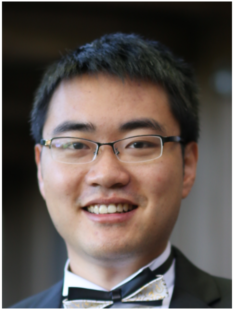
Cong Ma

Berkeley
UNIVERSITY OF CALIFORNIA

FODSI seminar, April 9^{th}, 2021



Banghua Zhu
EECS

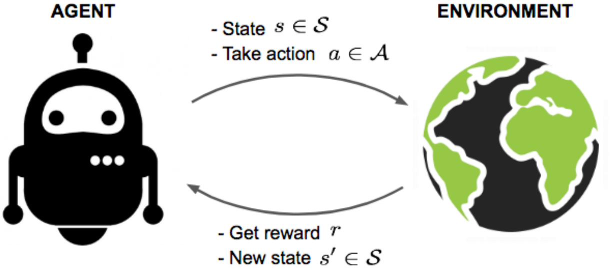


Jiantao Jiao
EECS & Stat

Martin Wainwright
EECS & Stat

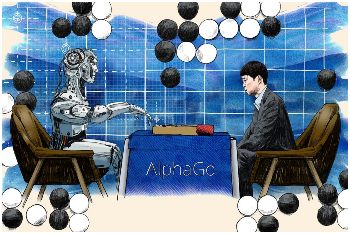
Berkeley
UNIVERSITY OF CALIFORNIA

Reinforcement learning (RL)

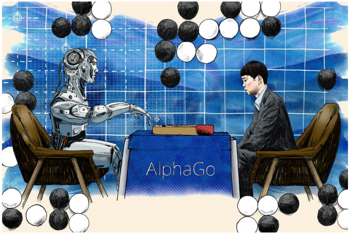


Goal: learn an optimal policy to maximize rewards

A key ingredient: policy evaluation



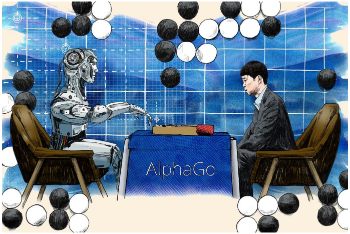
A key ingredient: policy evaluation



on-policy evaluation

deploy policy in environment

A key ingredient: policy evaluation

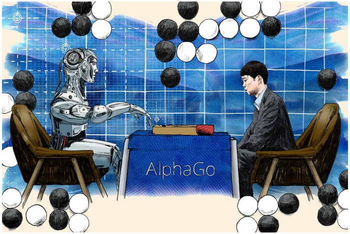


on-policy evaluation

deploy policy in environment

— *costly, dangerous, unethical*

A key ingredient: policy evaluation



on-policy evaluation

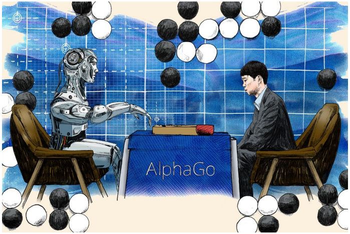
deploy policy in environment

— *costly, dangerous, unethical*

off-policy evaluation (OPE)

leverage historical data

A key ingredient: policy evaluation



on-policy evaluation

deploy policy in environment

— *costly, dangerous, unethical*

off-policy evaluation (OPE)

leverage historical data

— *distribution shift!*

This talk

Off-policy evaluation for multi-armed bandits

— *how to optimally tackle distribution shift*

This talk

Off-policy evaluation for multi-armed bandits

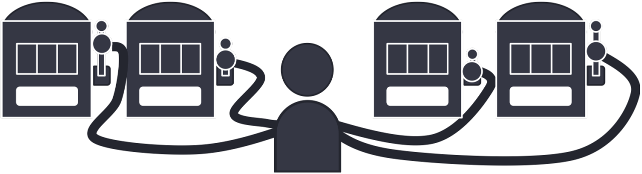
— *how to optimally tackle distribution shift*

"Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism"

— with P. Rashidinejad, B. Zhu, J. Jiao, and S. Russell

Background: multi-armed bandits and OPE

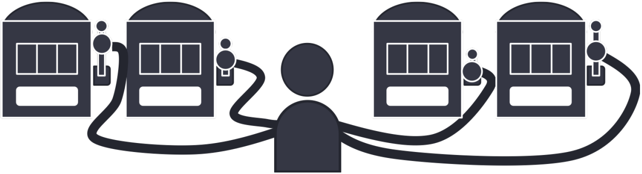
Multi-armed bandits



- Action space: $\mathcal{A} = [k] \coloneqq \{1, 2, \ldots, k\}$
- Reward distributions: $f \coloneqq \{f(\cdot \mid a)\}_{a \in \mathcal{A}}$

$$\mathcal{F}(r_{\max}) \coloneqq \{f \mid \operatorname{supp}(f(\cdot \mid a)) \subseteq [0, r_{\max}] \text{ for each } a \in [k]\}$$

Multi-armed bandits

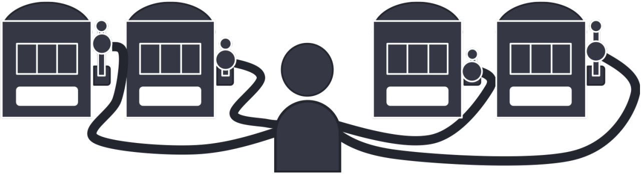


- Action space: $\mathcal{A} = [k] := \{1, 2, \ldots, k\}$
- Reward distributions: $f := \{f(\cdot \mid a)\}_{a \in \mathcal{A}}$

$$\mathcal{F}(r_{\max}) := \{f \mid \mathrm{supp}(f(\cdot \mid a)) \subseteq [0, r_{\max}] \text{ for each } a \in [k]\}$$

- Policy π: a distribution over $[k]$

Multi-armed bandits



- Action space: $\mathcal{A} = [k] \coloneqq \{1, 2, \ldots, k\}$
- Reward distributions: $f \coloneqq \{f(\cdot \mid a)\}_{a \in \mathcal{A}}$

$$\mathcal{F}(r_{\max}) \coloneqq \{f \mid \mathrm{supp}(f(\cdot \mid a)) \subseteq [0, r_{\max}] \text{ for each } a \in [k]\}$$

- Policy π: a distribution over $[k]$
- Value function of a policy: $V_f(\pi) \coloneqq \sum_{a \in [k]} \pi(a) r_f(a)$

— $r_f(a)$: mean reward of $f(\cdot \mid a)$

OPE in multi-armed bandits

Given

- observed data: $\{(A_i, R_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \pi_{\mathsf{b}} \otimes f$
- target policy π_{t}

Goal: estimate value function of target policy

$$V_f(\pi_{\mathsf{t}}) = \sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a)$$

Two classical estimators

Goal: estimate value function of target policy

$$V_f(\pi_{\mathsf{t}}) = \sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a)$$

plug-in estimator

$$\widehat{V}_{\mathsf{plug}} := \sum_{a \in [k]} \pi_{\mathsf{t}}(a) \widehat{r}(a)$$

$\widehat{r}(a) :=$ empirical mean reward

Two classical estimators

Goal: estimate value function of target policy

$$V_f(\pi_{\mathsf{t}}) = \sum_{a \in [k]} \pi_{\mathsf{t}}(a) r_f(a)$$

plug-in estimator

$$\widehat{V}_{\mathsf{plug}} := \sum_{a \in [k]} \pi_{\mathsf{t}}(a) \widehat{r}(a)$$

$\widehat{r}(a) :=$ empirical mean reward

importance sampling estimator

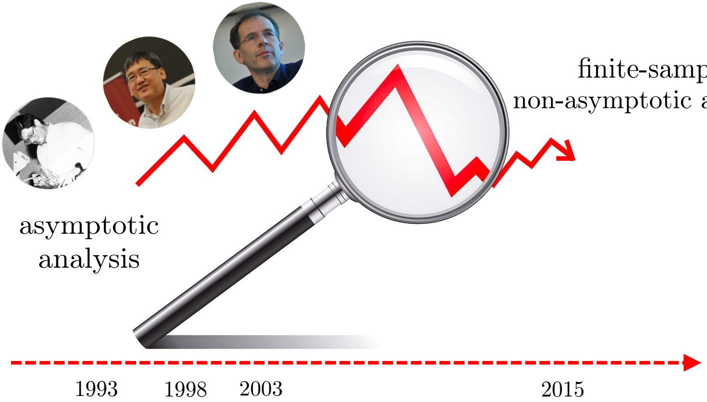
$$\widehat{V}_{\mathsf{IS}} := \frac{1}{n} \sum_{i \in [n]} \rho(A_i) R_i$$

$$\rho(a) := \frac{\pi_{\mathsf{t}}(a)}{\pi_{\mathsf{b}}(a)}$$

Gaps in statistical understanding of OPE

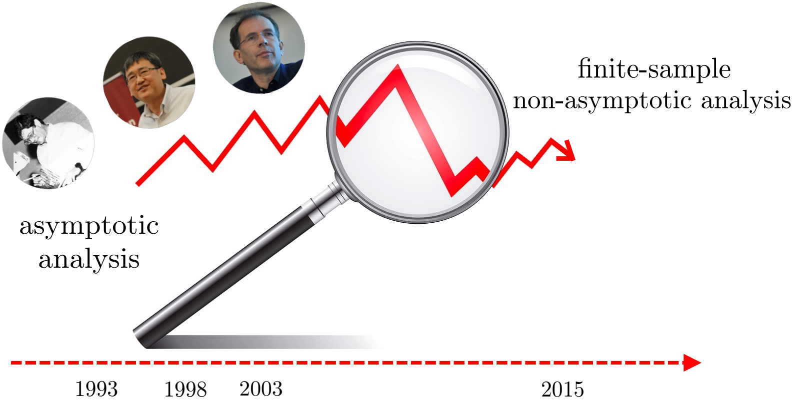
— a few motivating questions

Non-asymptotic analysis of OPE



— *notable exceptions: Li et al. '15, Wang et al. '17*

Non-asymptotic analysis of OPE



— *notable exceptions: Li et al. '15, Wang et al. '17*

Can we develop procedures that are optimal for all sample sizes?

Known vs. unknown behavior policies

Known behavior policy

VS

Unknown behavior policy

Known vs. unknown behavior policies

Known behavior policy

VS

Unknown behavior policy

Is there statistical difference between knowing and not knowing π_{b}?

Known vs. unknown behavior policies

Known behavior policy

VS

Unknown behavior policy

Is there statistical difference between knowing and not knowing π_{b}?

Asymptotic:
NO

Non-asymptotic
???

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

What if we have partial knowledge of behavior policy,

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

What if we have partial knowledge of behavior policy, say

- we know how close behavior policy is to target policy

$$\max_a \pi_{\mathsf{t}}(a)/\pi_{\mathsf{b}}(a) \leq U$$

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

What if we have partial knowledge of behavior policy, say

- we know how close behavior policy is to target policy

$$\max_a \pi_{\mathsf{t}}(a)/\pi_{\mathsf{b}}(a) \leq U$$

- or how well behavior policy explores action space

$$\min_a \pi_{\mathsf{b}}(a) \geq \nu$$

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

What if we have partial knowledge of behavior policy, say

- we know how close behavior policy is to target policy

$$\max_a \pi_{\mathsf{t}}(a)/\pi_{\mathsf{b}}(a) \leq U$$

- or how well behavior policy explores action space

$$\min_a \pi_{\mathsf{b}}(a) \geq \nu$$

Can we fully utilize such partial knowledge in OPE?

OPE with known behavior policy

Plug-in and importance sampling estimators

Goal: estimate value function of target policy

$$V_f(\pi_t) = \sum_{a \in [k]} \pi_t(a) r_f(a)$$

plug-in estimator

$$\widehat{V}_{\mathsf{plug}} := \sum_{a \in [k]} \pi_t(a) \widehat{r}(a)$$

$\widehat{r}(a) :=$ empirical mean reward

importance sampling estimator

$$\widehat{V}_{\mathsf{IS}} := \frac{1}{n} \sum_{i \in [n]} \rho(A_i) R_i$$

$$\rho(a) := \frac{\pi_t(a)}{\pi_b(a)}$$

Switch estimators

— *inspired by Wang et al. '17*

Switch estimators: for any subset $S \subseteq [k]$, we define

$$\widehat{V}_{\mathsf{switch}}(S) := \sum_{a \in S} \pi_{\mathsf{t}}(a)\widehat{r}(a) + \frac{1}{n}\sum_{i=1}^{n} \rho(A_i) R_i \mathbb{1}\{A_i \notin S\}$$

Switch estimators

— *inspired by Wang et al. '17*

Switch estimators: for any subset $S \subseteq [k]$, we define

$$\widehat{V}_{\mathsf{switch}}(S) := \sum_{a \in S} \pi_{\mathsf{t}}(a)\widehat{r}(a) + \frac{1}{n}\sum_{i=1}^{n} \rho(A_i) R_i \mathbb{1}\{A_i \notin S\}$$

- when $S = [k]$, recover plug-in estimator
- when $S = \emptyset$, recover importance sampling (IS) estimator
- Intermediate choices of S lead to interpolation between plug-in and IS estimators

Performance of Switch estimators

Proposition 1

For any subset $S \subseteq [k]$, *we have*

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] \leq 3r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

Performance of Switch estimators

Proposition 1

For any subset $S \subseteq [k]$, *we have*

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] \leq 3r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

— *How to choose subset* S*?*

Performance of Switch estimators

Proposition 1

For any subset $S \subseteq [k]$, we have

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] \leq 3r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

— *How to choose subset S?*

A simple idea:

$$\min_{S \subseteq [k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

Performance of Switch estimators

Proposition 1

For any subset $S \subseteq [k]$*, we have*

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S) - V_f(\pi_{\mathsf{t}}))^2] \le 3r_{\max}^2 \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

— *How to choose subset* S*?*

A simple idea:

$$\min_{S \subseteq [k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a \notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

— *combinatorial optimization problem!*

We can "solve" it!

Key convex program:

$$v^\star \in \arg\min_{\boldsymbol{v}\in\mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a\in[k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a\in[k]} |v(a)| \right\}$$

$$S^\star := \{a \mid v^\star(a) \neq 0\}$$

We can "solve" it!

Key convex program:

$$v^\star \in \arg\min_{\boldsymbol{v}\in\mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n}\sum_{a\in[k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]} |v(a)| \right\}$$

$$S^\star := \{a \mid v^\star(a) \neq 0\}$$

Proposition 2

$$\min_{S\subseteq[k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a\notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\} \asymp \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a\notin S^\star} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

We can "solve" it!

Key convex program:

$$v^\star \in \arg\min_{\boldsymbol{v}\in\mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n}\sum_{a\in[k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2}\sum_{a\in[k]} |v(a)| \right\}$$

$$S^\star := \{a \mid v^\star(a) \neq 0\}$$

Proposition 2

$$\min_{S\subseteq[k]} \left\{ \pi_{\mathsf{t}}^2(S) + \frac{\sum_{a\notin S} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\} \asymp \left\{ \pi_{\mathsf{t}}^2(S^\star) + \frac{\sum_{a\notin S^\star} \pi_{\mathsf{b}}(a)\rho^2(a)}{n} \right\}$$

— $\widehat{V}_{\mathsf{switch}}(S^\star)$ *is optimal among family of Switch estimators*

Is Switch estimator universally optimal?

Minimax risk of OPE:

$$\mathcal{R}_n^\star(\pi_\mathsf{t};\pi_\mathsf{b}) \coloneqq \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_\mathsf{b}\otimes f}[(\widehat{V} - V_f(\pi_\mathsf{t}))^2]$$

Is Switch estimator universally optimal?

Minimax risk of OPE:

$$\mathcal{R}_n^\star(\pi_\mathsf{t};\pi_\mathsf{b}) \coloneqq \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_\mathsf{b}\otimes f}[(\widehat{V} - V_f(\pi_\mathsf{t}))^2]$$

Theorem 1

For all pairs $(\pi_\mathsf{b}, \pi_\mathsf{t})$ and for all n, we have

$$\mathcal{R}_n^\star(\pi_\mathsf{t};\pi_\mathsf{b}) \gtrsim r_{\max}^2 \left\{ \pi_\mathsf{t}^2(S^\star) + \frac{\sum_{a\notin S^\star} \pi_\mathsf{b}(a)\rho^2(a)}{n} \right\}$$

— *Switch estimator is minimax optimal for all sample sizes*

Sanity checks

- Degenerate case of on-policy evaluation, i.e., $\pi_{\mathsf{t}} = \pi_{\mathsf{b}}$
  We know IS estimator (a.k.a. Monte Carlo estimator) is optimal

$$\widehat{V}_{\mathsf{IS}} = \frac{1}{n}\sum_{i=1}^{n} \rho(A_i) R_i = \frac{1}{n}\sum_{i=1}^{n} R_i$$

  with optimal rate $r_{\max}^2/n$

Sanity checks

- Degenerate case of on-policy evaluation, i.e., $\pi_{\mathsf{t}} = \pi_{\mathsf{b}}$
  We know IS estimator (a.k.a. Monte Carlo estimator) is optimal

  $$\widehat{V}_{\mathsf{IS}} = \frac{1}{n}\sum_{i=1}^{n} \rho(A_i) R_i = \frac{1}{n}\sum_{i=1}^{n} R_i$$

  with optimal rate $r_{\max}^2/n$

It can be shown from our minimax theorem that $S^\star = \emptyset$ in this case

Sanity checks

- Large-sample regime: in general when $\pi_{\mathsf{t}} \neq \pi_{\mathsf{b}}$, one can show that when

$$n \gg \frac{\max_{a \in [k]} \rho^2(a)}{\sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a)},$$

  $S^\star = \emptyset$, and hence IS estimator is optimal, with rate $r_{\max}^2 \cdot \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a)/n$

Sanity checks

- Large-sample regime: in general when $\pi_{\mathsf{t}} \neq \pi_{\mathsf{b}}$, one can show that when

$$n \gg \frac{\max_{a \in [k]} \rho^2(a)}{\sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a)},$$

  $S^\star = \emptyset$, and hence IS estimator is optimal, with rate $r_{\max}^2 \cdot \sum_{a \in [k]} \pi_{\mathsf{b}}(a) \rho^2(a) / n$

* recover large-sample result in Li et al. '15 (bounded reward setting)
* our results accommodate any sample size, especially small sample size where IS could perform poorly

Numerics

Setup: $\pi_{\mathsf{t}}(a) = 1/k$, $f(\cdot \mid a) = \mathrm{Bern}(0.5)$ for all $a \in [k]$, $n = 1.5k$

$$\pi_{\mathsf{b}}(1) = \pi_{\mathsf{b}}(2) = \cdots = \pi_{\mathsf{b}}(\sqrt{k}) = \frac{1}{k^2},$$

$$\pi_{\mathsf{b}}(\sqrt{k}+1) = \pi_{\mathsf{b}}(\sqrt{k}+2) = \cdots = \pi_{\mathsf{b}}(k) = \frac{1 - \frac{1}{k^{3/2}}}{k - \sqrt{k}}$$

Numerics

Setup: $\pi_{\mathsf{t}}(a) = 1/k,\ f(\cdot \mid a) = \mathrm{Bern}(0.5)$ for all $a \in [k]$, $n = 1.5k$

$$\pi_{\mathsf{b}}(1) = \pi_{\mathsf{b}}(2) = \cdots = \pi_{\mathsf{b}}(\sqrt{k}) = \frac{1}{k^2},$$

$$\pi_{\mathsf{b}}(\sqrt{k}+1) = \pi_{\mathsf{b}}(\sqrt{k}+2) = \cdots = \pi_{\mathsf{b}}(k) = \frac{1 - \frac{1}{k^{3/2}}}{k - \sqrt{k}}$$

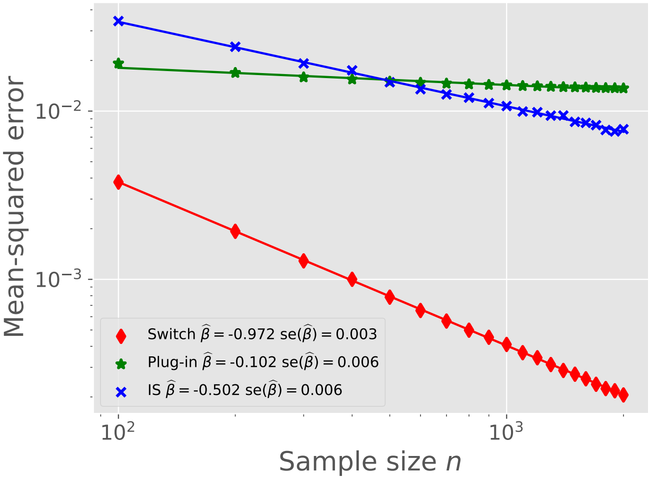
Theoretical predictions:

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2] \asymp 1,$$

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{IS}} - V_f(\pi_{\mathsf{t}}))^2] \asymp n^{-1/2}, \quad \text{and}$$

$$\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{switch}}(S^\star) - V_f(\pi_{\mathsf{t}}))^2] \asymp n^{-1}$$

Numerics (cont.)



A closer look at Switch estimator

Switch estimator:

$$\widehat{V}_{\mathsf{switch}}(S^\star) := \sum_{a \in S^\star} \pi_{\mathsf{t}}(a)\widehat{r}(a) + \frac{1}{n}\sum_{i=1}^{n} \rho(A_i) R_i \mathbb{1}\{A_i \notin S^\star\}$$

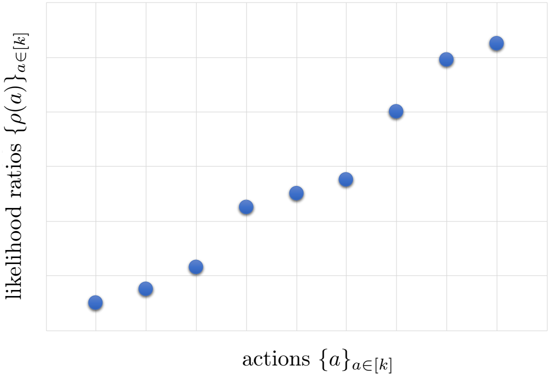
Key convex program:

$$v^\star \in \arg\min_{\boldsymbol{v} \in \mathbb{R}^k} \left\{ \sqrt{\frac{1}{8n} \sum_{a \in [k]} \frac{[\pi_{\mathsf{t}}(a) - v(a)]^2}{\pi_{\mathsf{b}}(a)}} + \frac{1}{2} \sum_{a \in [k]} |v(a)| \right\}$$

$$S^\star := \{a \mid v^\star(a) \neq 0\}$$

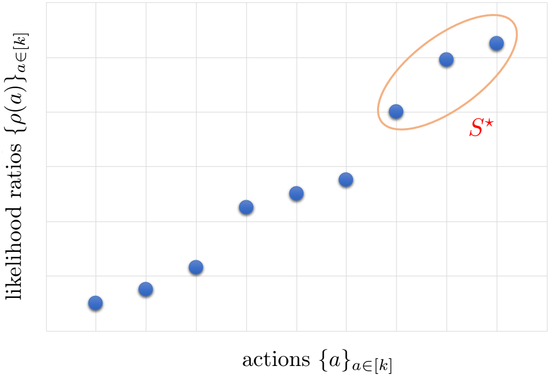
A closer look at Switch estimator

Without loss of generality, we assume



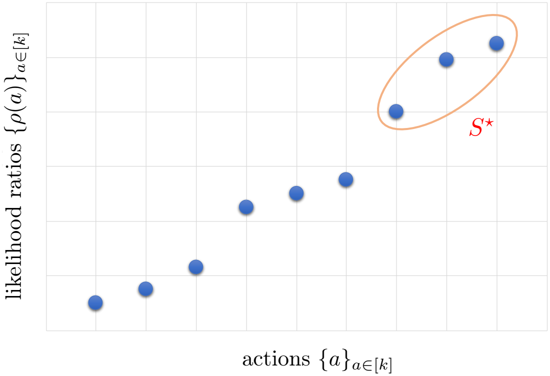
A closer look at Switch estimator

$S^\star$—if nonempty—must contain actions with largest likelihood ratios



A closer look at Switch estimator

$S^\star$—if nonempty—must contain actions with largest likelihood ratios



Key message: Switch optimally truncates large likelihood ratios

— *variance reduction*

OPE with unknown behavior policy

What's the right performance metric?

- First attempt: global worst-case risk of $\widehat{V}$

$$\sup_{\pi_{\mathsf{b}}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f} \big[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2 \big]$$

What's the right performance metric?

- First attempt: global worst-case risk of $\widehat{V}$

$$\sup_{\pi_{\mathsf{b}}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]$$

- Failure of attempt:

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2] \asymp r_{\max}^2$$

in addition, $\widehat{V} \equiv 0$ is minimax optimal...

What's the right performance metric?

- First attempt: global worst-case risk of $\widehat{V}$

$$\sup_{\pi_{\mathsf{b}}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f} [(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]$$

- Failure of attempt:

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f} [(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2] \asymp r_{\max}^2$$

  in addition, $\widehat{V} \equiv 0$ is minimax optimal...

- Rationale for failure: adversary can choose bad behavior policy without paying price

Competitive ratio

— *inspired by online learning literature*

Worst-case competitive ratio of $\widehat{V}$:

$$\mathcal{C}(\widehat{V};\pi_{\mathsf{t}}) \coloneqq \sup_{\pi_{\mathsf{b}}, f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})}$$

— $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) = \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]$

Competitive ratio

— *inspired by online learning literature*

Worst-case competitive ratio of $\widehat{V}$:

$$\mathcal{C}(\widehat{V};\pi_{\mathsf{t}}) \coloneqq \sup_{\pi_{\mathsf{b}}, f\in\mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}})}$$

— $\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{b}}) = \inf_{\widehat{V}} \sup_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]$

- Proof of concept: when $\widehat{V} \equiv 0$, we have

$$\mathcal{C}(\widehat{V};\pi_{\mathsf{t}}) \geq \frac{\mathbb{E}_{\pi_{\mathsf{t}}\otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}};\pi_{\mathsf{t}})} \asymp \frac{(V_f(\pi))^2}{r_{\max}^2/n} \asymp n$$

Competitive ratio of plug-in estimator

Theorem 2

For any target policy π_{t}, plug-in estimator $\widehat{V}_{\mathsf{plug}}$ satisfies

$$\sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \lesssim |\operatorname{supp}(\pi_{\mathsf{t}})|$$

Competitive ratio of plug-in estimator

Theorem 2

For any target policy π_{t}, plug-in estimator $\widehat{V}_{\mathsf{plug}}$ satisfies

$$\sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \lesssim |\operatorname{supp}(\pi_{\mathsf{t}})|$$

- Worst-case competitive ratio is at most k (since $|\operatorname{supp}(\pi_{\mathsf{t}})| \leq k$) $\Longrightarrow$ plug-in estimator is strictly better than all-zeros estimator

Competitive ratio of plug-in estimator

Theorem 2

For any target policy π_{t}, plug-in estimator $\widehat{V}_{\mathsf{plug}}$ satisfies

$$\sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V}_{\mathsf{plug}} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \lesssim |\operatorname{supp}(\pi_{\mathsf{t}})|$$

- Worst-case competitive ratio is at most k (since $|\operatorname{supp}(\pi_{\mathsf{t}})| \leq k$) $\Longrightarrow$ plug-in estimator is strictly better than all-zeros estimator
- Adaptivity of plug-in estimator to target policy

Is plug-in estimator optimal?

Theorem 3

Suppose that sample size obeys $n \gg \frac{k}{\log k}$. *Then for each* $s \in \{1, 2, \dots, k\}$, *there exists a target policy* π_t *supported on* s *actions and*

$$\inf_{\widehat{V}} \sup_{\pi_\mathsf{b}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_\mathsf{b} \otimes f}[(\widehat{V} - V_f(\pi_\mathsf{t}))^2]}{\mathcal{R}_n^\star(\pi_\mathsf{t}; \pi_\mathsf{b})} \gtrsim \max\left\{\frac{s}{\log k}, 1\right\}$$

Is plug-in estimator optimal?

Theorem 3

Suppose that sample size obeys $n \gg \frac{k}{\log k}$. *Then for each* $s \in \{1, 2, \dots, k\}$, *there exists a target policy* π_{t} *supported on* s *actions and*

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^{\star}(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \gtrsim \max\left\{\frac{s}{\log k}, 1\right\}$$

- Plug-in estimator is rate-optimal up to a log factor

Is plug-in estimator optimal?

Theorem 3

Suppose that sample size obeys $n \gg \frac{k}{\log k}$. *Then for each* $s \in \{1, 2, \ldots, k\}$, *there exists a target policy* π_{t} *supported on* s *actions and*

$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^{\star}(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \gtrsim \max\left\{\frac{s}{\log k}, 1\right\}$$

- Plug-in estimator is rate-optimal up to a log factor
- Performance difference between knowing and not knowing behavior policy scales as $|\operatorname{supp}(\pi_{\mathsf{t}})|$

Is plug-in estimator optimal?

Theorem 3

Suppose that sample size obeys $n \gg \frac{k}{\log k}$. *Then for each* $s \in \{1, 2, \dots, k\}$, *there exists a target policy* π_{t} *supported on* s *actions and*

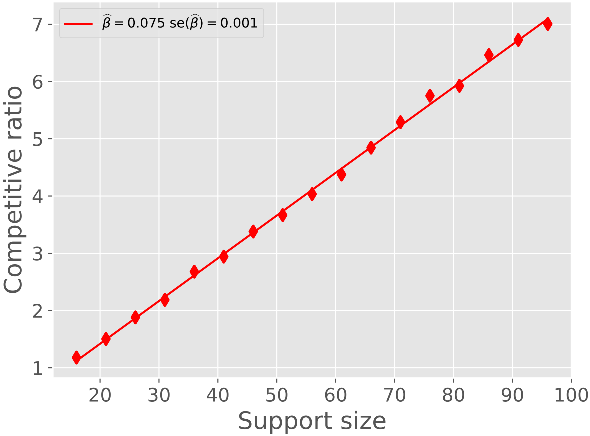
$$\inf_{\widehat{V}} \sup_{\pi_{\mathsf{b}}, f \in \mathcal{F}} \frac{\mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]}{\mathcal{R}_n^\star(\pi_{\mathsf{t}}; \pi_{\mathsf{b}})} \gtrsim \max\left\{\frac{s}{\log k}, 1\right\}$$

- Plug-in estimator is rate-optimal up to a log factor
- Performance difference between knowing and not knowing behavior policy scales as $|\mathrm{supp}(\pi_{\mathsf{t}})|$

— *in contrast to asymptotics*

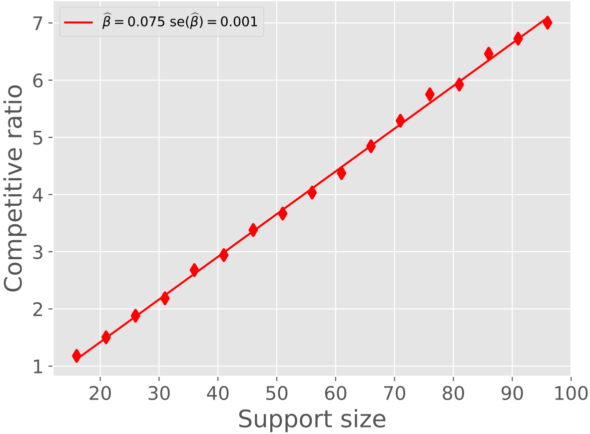
Numerics

Setup: $k = 100$, $n = 2k$, fix π_{b} and vary π_{t} uniform over $[s]$



Numerics

Setup: $k = 100$, $n = 2k$, fix π_{b} and vary π_{t} uniform over $[s]$



Knowing behavior policy helps in non-asymptotics!

OPE with partial knowledge of behavior policy

OPE with partial knowledge of behavior policy

Known behavior policy

Unknown behavior policy

What if we have partial knowledge of behavior policy?

Our focus: minimum exploration probability

$$\pi_{\mathsf{b}} \in \Pi(\nu) := \{\pi \mid \min_{a \in [k]} \pi(a) \geq \nu\}$$

— $\nu \in [0, 1/k]$

Optimal estimators

Goal: develop estimators that can achieve

$$\inf_{\widehat{V}} \sup_{(\pi_{\mathsf{b}}, f) \in \Pi(\nu) \times \mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}} \otimes f}[(\widehat{V} - V_f(\pi_{\mathsf{t}}))^2]$$

Optimal estimators

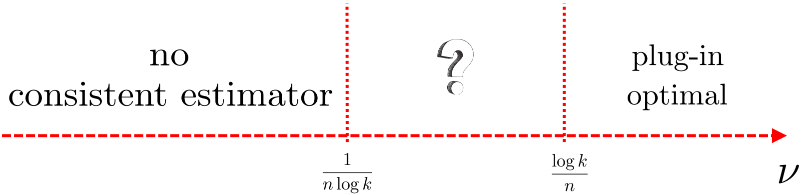
Goal: develop estimators that can achieve

$$\inf_{\widehat{V}} \sup_{(\pi_{\mathsf{b}},f)\in\Pi(\nu)\times\mathcal{F}} \mathbb{E}_{\pi_{\mathsf{b}}\otimes f}[(\widehat{V}-V_f(\pi_{\mathsf{t}}))^2]$$

no consistent estimator	?	plug-in optimal

$\frac{1}{n\log k}$ $\quad$ $\frac{\log k}{n}$ $\quad$ ν

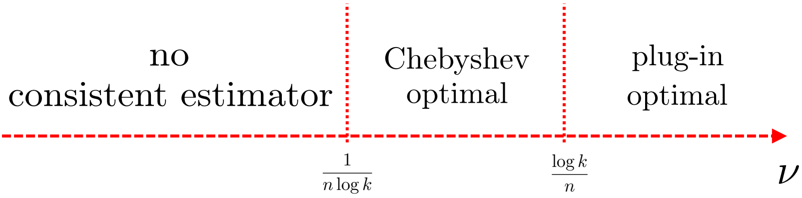
Tackling less-exploratory π_b



- Why plug-in fails when π_b is less exploratory?

  — large bias due to insufficient observations

Tackling less-exploratory π_b



- Why plug-in fails when π_b is less exploratory?

  — large bias due to insufficient observations

- How to reduce bias?

  — draw connection to support size estimation (cf. Wu and Yang '16)

  — best polynomial approximation

A peek at Chebyshev estimator

plug-in estimator

$$\widehat{V}_{\mathsf{plug}} := \sum_{a \in [k]} \pi_{\mathsf{t}}(a)\widehat{r}(a)$$

$\widehat{r}(a) :=$ empirical mean reward

Chebyshev estimator

$$\widehat{V}_{\mathsf{C}} := \sum_{a \in [k]} \pi_{\mathsf{t}}(a)\widehat{r}(a)g_L(n(a))$$

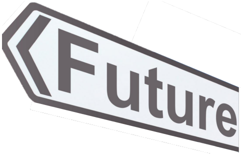
$g_L(n(a)) :=$ Chebyshev poly

Concluding remarks

- Known π_{b}: Switch is minimax optimal for all sample sizes
- Unknown π_{b}: fundamentally different, plug-in is near-optimal
- Partial knowledge: improvement is possible, bias reduction is needed

Concluding remarks

- Known π_{b}: Switch is minimax optimal for all sample sizes
- Unknown π_{b}: fundamentally different, plug-in is near-optimal
- Partial knowledge: improvement is possible, bias reduction is needed



- Extension to other reward families
- Smooth characterization of gap between knowing and not knowing π_{b}
- Adaptivity to $\min_a \pi_{\mathsf{b}}(a)$

Paper:

"Minimax Off-Policy Evaluation for Multi-Armed Bandits,"

C. Ma, B. Zhu, J. Jiao, M. J. Wainwright, arXiv:2101.07781, 2021