Deeply-Debiased Off-Policy Interval Estimation

Chengchun Shi*[1] **Runzhe Wan***[2] Victor Chernozhukov[3]
Rui Song[2]

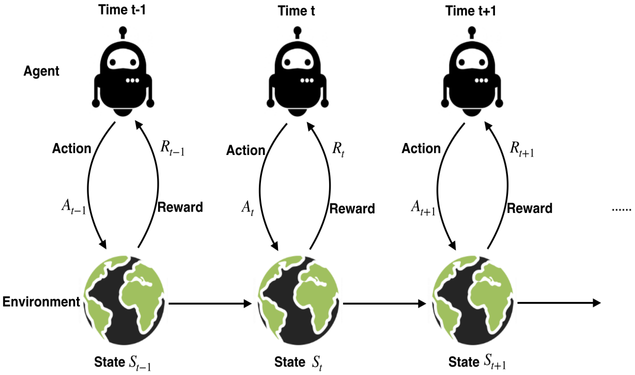
[1]London School of Economics and Political Science

[2]North Carolina State University

[3]Massachusetts Institute of Technology

ICML 2021

Reinforcement Learning



Objective: find an optimal policy that maximizes the cumulative reward.

Off-policy Evaluation

In many **real-world** applications, a direct deployment of RL policies can be **costly, risky, unethical, or even infeasible**.

(a) Finance

(b) Mobile health

(c) Autonomous driving

Off-policy Evaluation

In many **real-world** applications, a direct deployment of RL policies can be **costly, risky, unethical, or even infeasible**.

(a) Finance

(b) Mobile health

(c) Autonomous driving

Off-Policy Evaluation (OPE): Using historical data generated from a *behavior policy* b to evaluate the impact of a different *target policy* π

$$\eta^{\pi} = \mathbb{E}_{s \sim \mathbb{G}} V^{\pi}(s),$$

for a given reference distribution $\mathbb{G}$.

Off-policy Evaluation

In many **real-world** applications, a direct deployment of RL policies can be **costly, risky, unethical, or even infeasible**.

(a) Finance

(b) Mobile health

(c) Autonomous driving

Off-Policy Evaluation (OPE): Using historical data generated from a *behavior policy* b to evaluate the impact of a different *target policy* π

$$\eta^{\pi} = \mathbb{E}_{s \sim \mathbb{G}} V^{\pi}(s),$$

for a given reference distribution $\mathbb{G}$.

Confidence interval (CI): For high-stake applications, in addition to a point estimate, it is crucial to construct a CI that quantifies its uncertainty.

Existing Methods

Existing Methods

Point estimator

- Direct method: model-free [LVY19, YW20, DW20] or model-based [JL16, TB16, HSN17]
- Importance sampling: [T15, JL16, LLT18]
- Doubly robust: [FCG18, UHJ19, TB16] , including the state-of-the-art DRL [KU19]

Existing Methods

Point estimator

- Direct method: model-free [LVY19, YW20, DW20] or model-based [JL16, TB16, HSN17]
- Importance sampling: [T15, JL16, LLT18]
- Doubly robust: [FCG18, UHJ19, TB16] , including the state-of-the-art DRL [KU19]

CI estimator

- Much less studied
- Bootstrapping [TTG15, HSN17, KN20, HDL21]: computationally heavy
- Concentration inequality [FRT20, TTG15]: not tight
- Empirical likelihood [DNC20]: may not valid under serial dependence

Existing Methods

Point estimator

- Direct method: model-free [LVY19, YW20, DW20] or model-based [JL16, TB16, HSN17]
- Importance sampling: [T15, JL16, LLT18]
- Doubly robust: [FCG18, UHJ19, TB16] , including the state-of-the-art DRL [KU19]

CI estimator

- Much less studied
- Bootstrapping [TTG15, HSN17, KN20, HDL21]: computationally heavy
- Concentration inequality [FRT20, TTG15]: not tight
- Empirical likelihood [DNC20]: may not valid under serial dependence

*Q: is it possible to develop a **robust** and **efficient** off-policy value estimator, with rigorous **uncertainty quantification** under **practically feasible conditions**?*

Advances of the Proposed Method: D2OPE

By leveraging the statistical efficiency of DRL and our novel deeply-debiasing procedure, D2OPE is:

Advances of the Proposed Method: D2OPE

By leveraging the statistical efficiency of DRL and our novel deeply-debiasing procedure, D2OPE is:

- **robust**: the value estimator is **triply robust to model misspecifications** of the nuisance functions;

Advances of the Proposed Method: D2OPE

By leveraging the statistical efficiency of DRL and our novel deeply-debiasing procedure, D2OPE is:

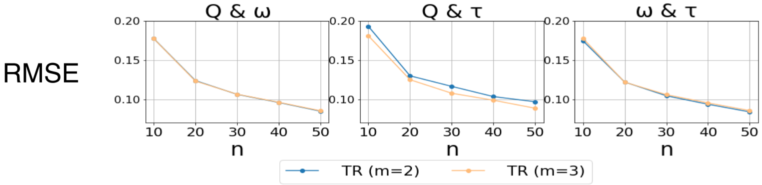
- **robust**: the value estimator is **triply robust to model misspecifications** of the nuisance functions;
- **efficient**:
  - the value estimator is **semiparametric efficient**
  - the CI is **tight**

Advances of the Proposed Method: D2OPE

By leveraging the statistical efficiency of DRL and our novel deeply-debiasing procedure, D2OPE is:

- **robust**: the value estimator is **triply robust to model misspecifications** of the nuisance functions;
- **efficient**:
  - the value estimator is **semiparametric efficient**
  - the CI is **tight**
- **flexible**:
  - DRL-based CI may fail when the nuisance functions converge slower than $(NT)^{-1/2}$
  - our **CI is valid** under much weaker and practically more feasible conditions, which allow the Q- and marginalized density ratio-estimator to converge at **arbitrary slow rates**.

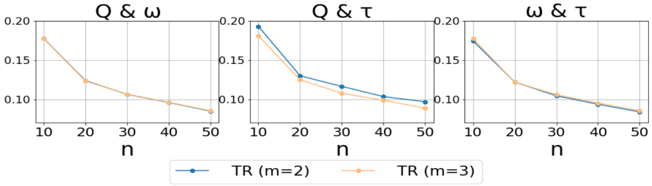
Toy Examples



(a) Our estimator is **consistent** as long as one of the three nuisance function estimators is.

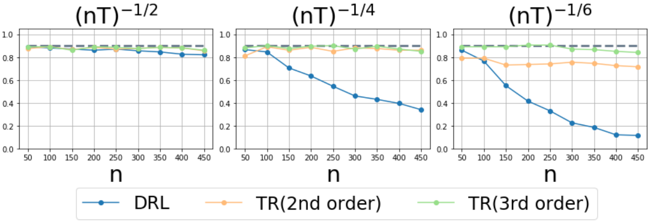
Toy Examples

RMSE



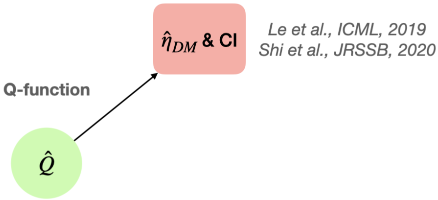
(a) Our estimator is **consistent** as long as one of the three nuisance function estimators is.

Coverage probabilities



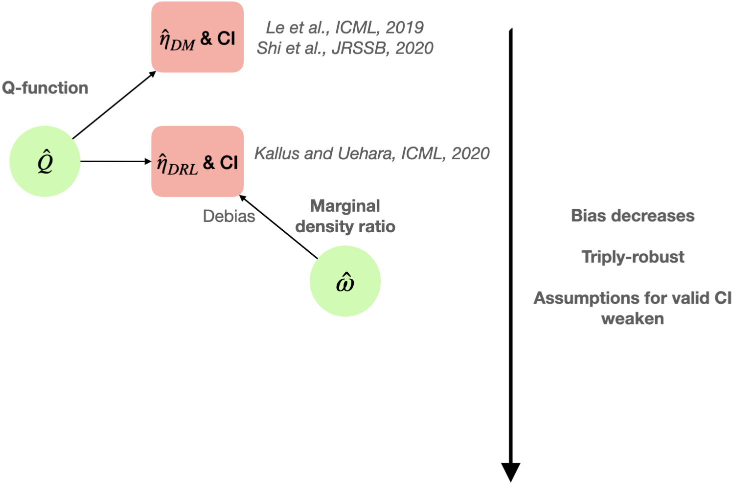
(b) Our CI is **valid** even when the nuisance functions converge at a slow rate, while DRL fails.

Main Idea



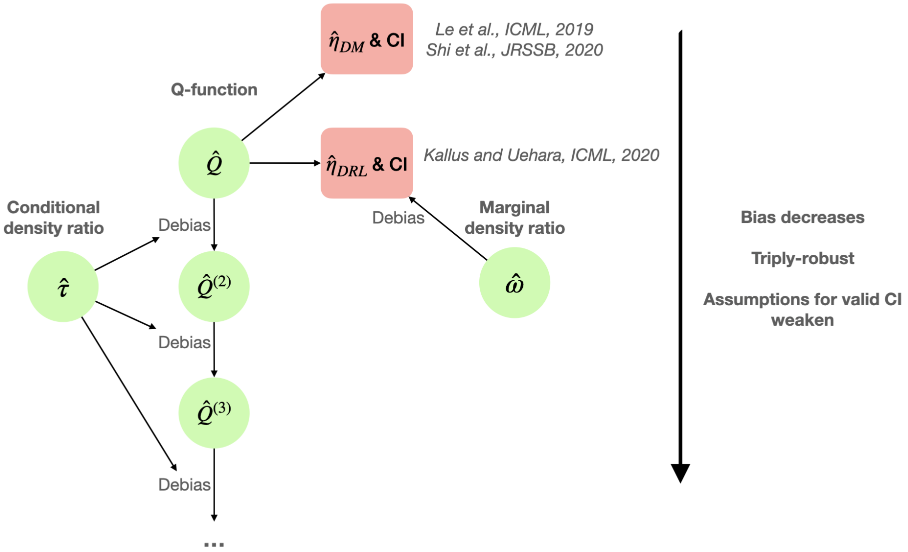
$\hat{\eta}_{DM}$ & CI
Le et al., ICML, 2019
Shi et al., JRSSB, 2020
Q-function
$\hat{Q}$

Main Idea



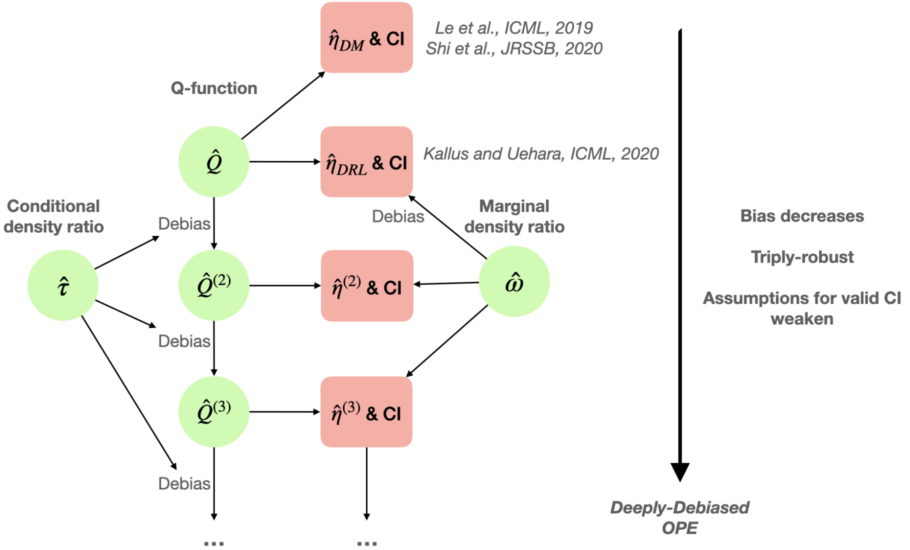
$\hat{\eta}_{DM}$ & CI
Le et al., ICML, 2019
Shi et al., JRSSB, 2020
Q-function
$\hat{Q}$
$\hat{\eta}_{DRL}$ & CI
Kallus and Uehara, ICML, 2020
Debias
Marginal density ratio
$\hat{\omega}$
Bias decreases
Triply-robust
Assumptions for valid CI weaken

Main Idea



$\hat{\eta}_{DM}$ & CI
Le et al., ICML, 2019
Shi et al., JRSSB, 2020
Q-function
$\hat{Q}$
$\hat{\eta}_{DRL}$ & CI
Kallus and Uehara, ICML, 2020
Conditional density ratio
Debias
Debias
Marginal density ratio
$\hat{\tau}$
$\hat{Q}^{(2)}$
$\hat{\omega}$
Debias
$\hat{Q}^{(3)}$
Debias
...
Bias decreases
Triply-robust
Assumptions for valid CI weaken

Main Idea



$\hat{\eta}_{DM}$ & CI
Le et al., ICML, 2019
Shi et al., JRSSB, 2020
Q-function
$\hat{Q}$
$\hat{\eta}_{DRL}$ & CI
Kallus and Uehara, ICML, 2020
Conditional density ratio
Debias
Debias
Marginal density ratio
$\hat{\tau}$
$\hat{Q}^{(2)}$
$\hat{\eta}^{(2)}$ & CI
$\hat{\omega}$
Debias
$\hat{Q}^{(3)}$
$\hat{\eta}^{(3)}$ & CI
Debias
...
...
Bias decreases
Triply-robust
Assumptions for valid CI weaken
Deeply-Debiased OPE

D2OPE: Deeply-Debiased OPE

- **Step 1. Data Splitting:** value estimation with each fold is based on nuisance function learned from the other folds;
- **Step 2. Estimation of nuisance functions:** learn three nuisance functions Q, ω, τ as $\widehat{Q}$, $\widehat{\omega}$, $\widehat{\tau}$;
- **Step 3. Debias Iteration:** iteratively debias $\widehat{Q}$ with $\widehat{\tau}$ for m times to obtain $\widehat{Q}^{(m+1)}$;
- **Step 4. Construction of the value estimator & CI:** constructing the m-th order estimator & its CI with $\widehat{Q}^{(m)}$ and $\widehat{\omega}$

Nuisance Functions

- Q-function: $Q^{\pi}(a,s) = \mathbb{E}^{\pi}(\sum_{t=0}^{+\infty} \gamma^t R_t | A_0 = a, S_0 = s)$

Nuisance Functions

- Q-function: $Q^\pi(a,s) = \mathbb{E}^\pi(\sum_{t=0}^{+\infty} \gamma^t R_t | A_0 = a, S_0 = s)$
- Marginalized density ratio ω^π: density ratio over (a,s)

$$\omega^\pi(a,s) = \frac{(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t p_t^\pi(a,s)}{p_\infty(a,s)}$$

Nuisance Functions

- Q-function: $Q^\pi(a, s) = \mathbb{E}^\pi(\sum_{t=0}^{+\infty} \gamma^t R_t | A_0 = a, S_0 = s)$
- Marginalized density ratio ω^π: density ratio over (a, s)

$$\omega^\pi(a, s) = \frac{(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t p_t^\pi(a, s)}{p_\infty(a, s)}$$

- **Conditional density ratio** τ^π: density ratio over (a, s) starting from (a_0, s_0)

$$\tau^\pi(a, s, a_0, s_0) = \frac{(1-\gamma)\{\mathbb{I}(a = a_0, s = s_0) + \sum_{t=1}^{+\infty} \gamma^t p_t^\pi(a, s | a_0, s_0)\}}{p_\infty(a, s)},$$

Nuisance Functions

- Q-function: $Q^{\pi}(a,s) = \mathbb{E}^{\pi}(\sum_{t=0}^{+\infty} \gamma^t R_t | A_0 = a, S_0 = s)$
- Marginalized density ratio ω^{π}: density ratio over (a, s)

$$\omega^{\pi}(a,s) = \frac{(1-\gamma)\sum_{t=0}^{+\infty} \gamma^t p_t^{\pi}(a,s)}{p_{\infty}(a,s)}$$

- **Conditional density ratio** τ^{π}: density ratio over (a, s) starting from (a_0, s_0)

$$\tau^{\pi}(a,s,a_0,s_0) = \frac{(1-\gamma)\{\mathbb{I}(a=a_0, s=s_0) + \sum_{t=1}^{+\infty} \gamma^t p_t^{\pi}(a,s|a_0,s_0)\}}{p_{\infty}(a,s)},$$

Estimation: Q^{π} and ω^{π} can be learned by several algorithms in the literature [LLT18, LVY19, KU20]; τ^{π} can be learned by solving an optimization problem, based on a novel result established in this work.

Key Step: Debias iteration

- DRL: debias the *plug-in value estimator* $\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s)$ with the *marignalize density ratio* ω

$$\widehat{\eta}_{\mathrm{DRL}} = \mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\omega}(A_{i,t},S_{i,t}) \times\left\{R_{i,t} - \widehat{Q}(A_{i,t},S_{i,t}) + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}(a,S_{i,t+1})\right\}$$

Key Step: Debias iteration

- DRL: debias the *plug-in value estimator* $\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s)$ with the *marignalize density ratio* ω

$$\widehat{\eta}_{\mathrm{DRL}} = \mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\omega}(A_{i,t},S_{i,t}) \times\left\{R_{i,t} - \widehat{Q}(A_{i,t},S_{i,t}) + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}(a,S_{i,t+1})\right\}$$

- Our proposal: additionally debias *any Q-function* with the *conditional density ratio* τ

$$\widehat{Q}^{(m+1)}(a,s) = \widehat{Q}^{(m)}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\tau}(A_{i,t},S_{i,t},a,s) \times\left\{R_{i,t} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}^{(m)}(a',S_{i,t+1}) - \widehat{Q}^{(m)}(A_{i,t},S_{i,t})\right\},$$

Key Step: Debias iteration

- DRL: debias the *plug-in value estimator* $\mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s)$ with the *marignalize density ratio* ω

$$\widehat{\eta}_{\mathrm{DRL}} = \mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\omega}(A_{i,t},S_{i,t}) \times\left\{R_{i,t} - \widehat{Q}(A_{i,t},S_{i,t}) + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}(a,S_{i,t+1})\right\}$$

- Our proposal: additionally debias *any Q-function* with the *conditional density ratio* τ

$$\widehat{Q}^{(m+1)}(a,s) = \widehat{Q}^{(m)}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\tau}(A_{i,t},S_{i,t},a,s) \times\left\{R_{i,t} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}^{(m)}(a',S_{i,t+1}) - \widehat{Q}^{(m)}(A_{i,t},S_{i,t})\right\},$$

- Repeat the procedure iteratively can **deeply debias** the Q-function and our final value estimator.

Construction of Value Estimator and CI

***m*-th order value estimator** (with m-th order Q-function estimator):

$$\widehat{\eta}^{(m)} = \mathbb{E}_{s\sim\mathbb{G},a\sim\pi(\cdot|s)}\widehat{Q}^{(m)}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\omega}(A_{i,t},S_{i,t})$$
$$\times\left\{R_{i,t} - \widehat{Q}^{(m)}(A_{i,t},S_{i,t}) + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}^{(m)}(a,S_{i,t+1})\right\}$$

Construction of Value Estimator and CI

***m*-th order value estimator** (with m-th order Q-function estimator):

$$\widehat{\eta}^{(m)} = \mathbb{E}_{s\sim\mathbb{G}, a\sim\pi(\cdot|s)} \widehat{Q}^{(m)}(a,s) + \frac{1}{1-\gamma}(nT)^{-1}\sum_{i,t}\widehat{\omega}(A_{i,t}, S_{i,t})$$
$$\times\left\{R_{i,t} - \widehat{Q}^{(m)}(A_{i,t}, S_{i,t}) + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}^{(m)}(a, S_{i,t+1})\right\}$$

Wald-type CI:

$$[\widehat{\eta}^{(m)} - z_{\alpha/2}(nT)^{-1/2}\widehat{\sigma}^{(m)}, \widehat{\eta}^{(m)} + z_{\alpha/2}(nT)^{-1/2}\widehat{\sigma}^{(m)}]$$

Theoretical Guarantees

Under mild assumptions, the proposed value estimator and CI yield:

Theoretical Guarantees

Under mild assumptions, the proposed value estimator and CI yield:

Robustness

For any m, as either n or T diverges to infinity, the value estimator $\widehat{\eta}^{(m)}$ is **consistent** when **either one** of $\widehat{Q}_k$, $\widehat{\tau}_k$ or $\widehat{\omega}_k$ converges in L_2-norm to Q^π, τ^π or ω^π for any k.

Theoretical Guarantees

Under mild assumptions, the proposed value estimator and CI yield:

Robustness

For any m, as either n or T diverges to infinity, the value estimator $\widehat{\eta}^{(m)}$ is **consistent** when **either one** of $\widehat{Q}_k$, $\widehat{\tau}_k$ or $\widehat{\omega}_k$ converges in L_2-norm to Q^π, τ^π or ω^π for any k.

Efficiency

For any m, we have $\sqrt{nT}(\widehat{\eta}^{(m)} - \mathbb{E}\widehat{\eta}^{(m)}) \stackrel{d}{\to} N(0, \sigma^2)$ as either n or T approaches infinity, where σ^2 is the **semiparametric efficiency bound**.

Theoretical Guarantees

Under mild assumptions, the proposed value estimator and CI yield:

Robustness

For any m, as either n or T diverges to infinity, the value estimator $\widehat{\eta}^{(m)}$ is **consistent** when **either one** of $\widehat{Q}_k$, $\widehat{\tau}_k$ or $\widehat{\omega}_k$ converges in L_2-norm to Q^π, τ^π or ω^π for any k.

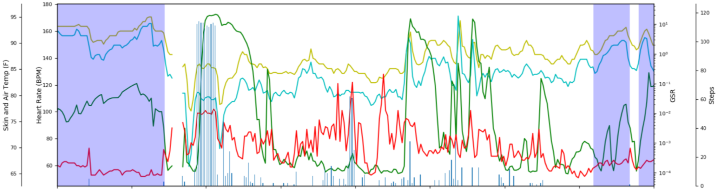
Efficiency

For any m, we have $\sqrt{nT}(\widehat{\eta}^{(m)} - \mathbb{E}\widehat{\eta}^{(m)}) \stackrel{d}{\to} N(0, \sigma^2)$ as either n or T approaches infinity, where σ^2 is the **semiparametric efficiency bound**.

Flexibility

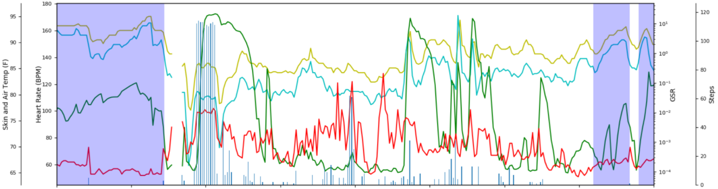
Suppose $\widehat{Q}$, $\widehat{\tau}$, and $\widehat{\omega}$ converge in L_2-norm at a rate of $(nT)^{-\alpha_1}$, $(nT)^{-\alpha_2}$, and $(nT)^{-\alpha_3}$, respectively. As long as the order m satisfies $\alpha_1 + (m-1)\alpha_2 + \alpha_3 > 1/2$, the proposed **CI achieves nominal coverage**.

Application: Mobile Health



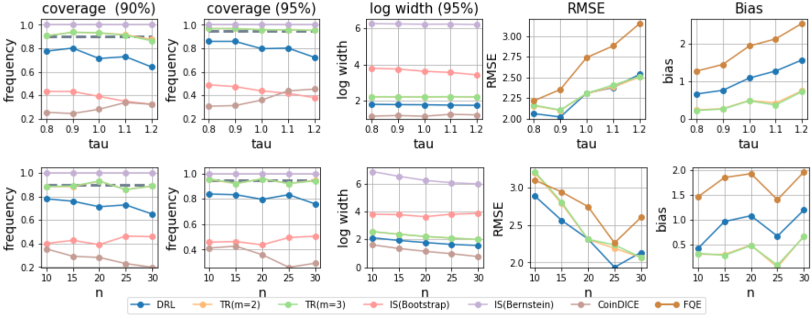
- OhioT1DM dataset [MB 18]: type-I diabetes
- State: patients' time-varying variables, e.g., glucose levels.
- Action: to inject (a certain amount of) insulin or not.
- Reward: the Index of Glycemic Control (RBJ09).

Application: Mobile Health



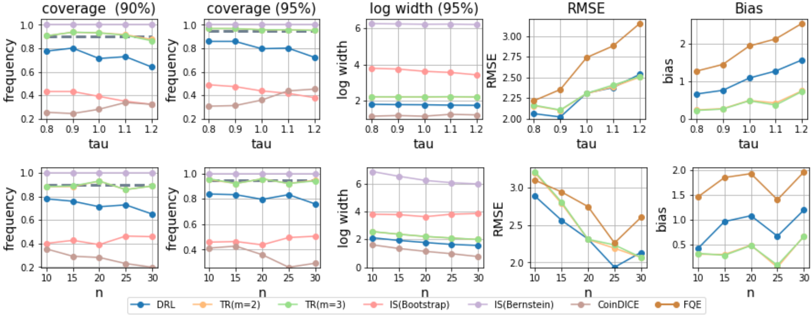
- OhioT1DM dataset [MB 18]: type-I diabetes
- State: patients' time-varying variables, e.g., glucose levels.
- Action: to inject (a certain amount of) insulin or not.
- Reward: the Index of Glycemic Control (RBJ09).
- **Objective: control the glucose level for patients with diabetes**

Result



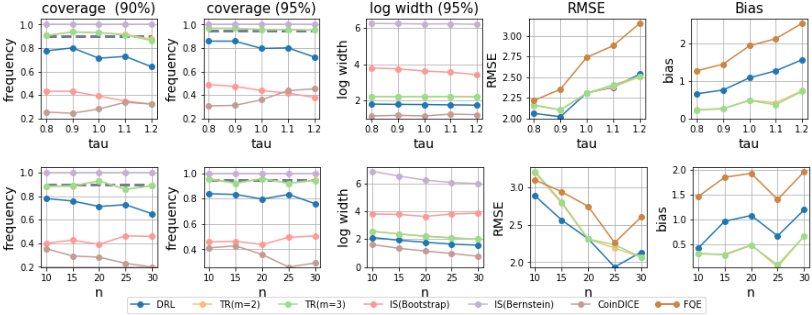
- Our CI is **valid**, while the other CIs may fail

Result



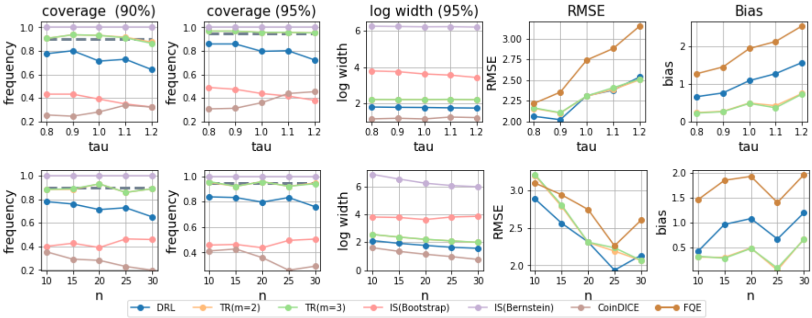
- Our CI is **valid**, while the other CIs may fail
- Our CI is **tight**, while IS-based suffer the curse of horizon

Result



- Our CI is **valid**, while the other CIs may fail
- Our CI is **tight**, while IS-based suffer the curse of horizon
- Our value estimator yields **low RMSE**, with **lower bias** than DRL

Result



- Our CI is **valid**, while the other CIs may fail
- Our CI is **tight**, while IS-based suffer the curse of horizon
- Our value estimator yields **low RMSE**, with **lower bias** than DRL
- **Similar results for CartPole** (OpenAI Gym)

Summary

- **D2OPE**: A provably robust, efficient and flexible OPE estimator with valid CI under practically feasible conditions.

Summary

- **D2OPE**: A provably robust, efficient and flexible OPE estimator with valid CI under practically feasible conditions.
- Preprint: https://arxiv.org/abs/2105.04646
- Code: https://github.com/RunzheStat/D2OPE

Thank you! ☺