Review of Off-Policy Evaluation (OPE)

Chengchun Shi

Outline

1. Off-Policy Evaluation (OPE) Introduction
2. OPE in Contextual Bandits
3. OPE in Reinforcement Learning

Outline

1. Off-Policy Evaluation (OPE) Introduction
2. OPE in Contextual Bandits
3. OPE in Reinforcement Learning

What is OPE and Why OPE

- **Objective**: Evaluate the impact of a **target policy** offline using historical data generated from a different **behavior policy**
- **Motivation**: In many applications, it can be **dangerous** to evaluate a **target policy** by directly running this policy

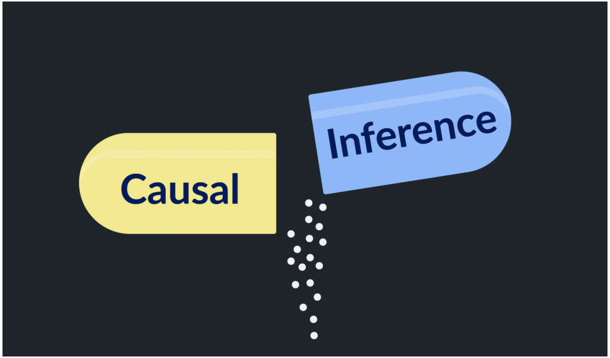
(a) Health Care

(b) Ridesharing

- **Healthcare**: which **medical treatment** to suggest for a patient
- **Ridesharing**: which **driver** to assign for a call order

Causal Inference

Off-policy evaluation is closely related to **causal inference**, whose objective is to learn the difference between a new treatment and a standard treatment



Outline

1. Off-Policy Evaluation (OPE) Introduction
2. **OPE in Contextual Bandits**
3. OPE in Reinforcement Learning

Contextual Bandits

- Extension of MAB with **contextual** information.
- A **widely-used** model in medicine and technological industries.
- At time $\boldsymbol{t}$, the agent
  - Observe a context $\boldsymbol{S_t}$;
  - Select an action $\boldsymbol{A_t}$;
  - Receives a reward $\boldsymbol{R_t}$ (depends on both $\boldsymbol{S_t}$ and $\boldsymbol{A_t}$).
- **Objective**: Given an i.i.d. offline dataset $\{(\boldsymbol{S_t}, \boldsymbol{A_t}, \boldsymbol{R_t}) : \boldsymbol{0} \le \boldsymbol{t} < \boldsymbol{T}\}$ generated by a behavior policy $\boldsymbol{b}$, i.e.,

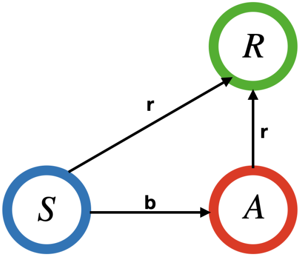
$$\Pr(\boldsymbol{A_t} = \boldsymbol{a} | \boldsymbol{S_t} = \boldsymbol{s}) = \boldsymbol{b}(\boldsymbol{a}|\boldsymbol{s}),$$

we aim to evaluate the mean outcome under a target policy $\boldsymbol{\pi}$, i.e.,

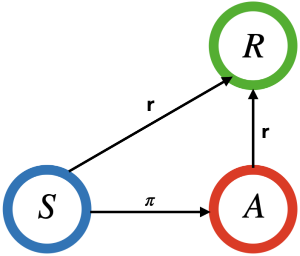
$$\Pr(\boldsymbol{A_t} = \boldsymbol{a} | \boldsymbol{S_t} = \boldsymbol{s}) = \boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s}).$$

Challenge

- **Confounding**: State serves as confounding variables that confound the action-reward pair
- **Distributional shift**: The target policy generally differs from the behavior policy



historical data



what we want to evaluate

Challenge (Cont'd)

- Suppose π is a nondynamic policy, i.e., there exists some a such that $\pi(a|s) = 1$ for any s. We aim to evaluate the value under a given action a. A naive estimator is

$$\frac{\sum_{t=0}^{T-1} R_t \mathbb{I}(A_t = a)}{\sum_{t=0}^{T-1} \mathbb{I}(A_t = a)} \xrightarrow{P} \mathbb{E}(R|A = a)$$

- This estimator is valid only when no confounding variables exist
- According to the causal diagram, the target policy's value equals

$$\mathbb{E}[\mathbb{E}(R|A = a, S)] \neq \mathbb{E}(R|A = a)$$

OPE Estimators

- With a general target policy π, the target policy's value equals

$$\sum_{a} \mathbb{E}[\pi(a|S)\mathbb{E}(R|A=a,S)] = \sum_{a} \mathbb{E}[\pi(a|S)r(S,a)],$$

  where $r(s,a) = \mathbb{E}(R|A=a, S=s)$
- Direct estimator
- Importance sampling estimator
- Doubly robust estimator

Direct Estimator

- Given that the target policy's value is given by

$$\sum_{a} \mathbb{E}[\pi(a|S)r(S,a)]$$

- The expectation can be approximated by the sample average, i.e.,

$$\frac{1}{T}\sum_{a}\sum_{t=0}^{T-1}[\pi(a|S_t)r(S_t,a)]$$

- The reward function can be replaced with some estimator $\widehat{r}$. This yields the direct estimator

$$\frac{1}{T}\sum_{a}\sum_{t=0}^{T-1}[\pi(a|S_t)\widehat{r}(S_t,a)]$$

Importance Sampling Estimator

- Given that the target policy's value is given by

$$\sum_{a} \mathbb{E}[\pi(a|S) r(S,a)]$$

- By the change of measure theory, it equals

$$\sum_{a} \mathbb{E}\left[b(a|S)\frac{\pi(a|S)}{b(a|S)} r(S,a)\right] = \mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)} r(S,A)\right] = \mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)} R\right]$$

- This yields the following importance sampling (IS) estimator [Zhang et al., 2012]

$$\frac{1}{T}\sum_{t=0}^{T-1} \frac{\pi(A_t|S_t)}{\widehat{b}(A_t|S_t)} R_t,$$

for a given estimator $\widehat{b}$

Direct Estimator v.s. IS Estimator

- Bias/Variance Trade-Off
- The direct estimator has **some bias**, since $\boldsymbol{r}$ needs to be estimated from data
- The IS estimator has **zero bias** when $\boldsymbol{b}$ is known as in randomized studies
- The IS estimator might have a **large variance** when $\boldsymbol{\pi}$ differs significantly from $\boldsymbol{b}$
- Suppose $R = r(S, A) + \varepsilon$ for some ε independent of (S, A),

$$\begin{aligned}\mathrm{Var}\left[\frac{\pi(A|S)}{b(A|S)}R\right] &= \mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)}\{R - r(S, A)\}\right]^2 + \text{some term}\\ &= \sigma^2\mathbb{E}\left[\frac{\pi^2(A|S)}{b^2(A|S)}\right] + \text{some term},\end{aligned}$$

where $\sigma^2 = \mathrm{Var}(\varepsilon)$

Extensions

- When π differs from b significantly, IS estimator suffers from **large variance** and becomes **unstable**
- Solutions sought by using **self-normalized** and/or **truncated** IS
- **Self-normalized** IS

$$\left[\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}\right]^{-1}\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_t$$

- **Truncated** IS

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{\max(\widehat{b}(A_t|S_t),\varepsilon)}R_t,$$

for some $\varepsilon > 0$

Doubly Robust Estimator

- Direct estimator

$$\frac{1}{T}\sum_{a}\sum_{t=0}^{T-1}[\pi(a|S_t)\widehat{r}(S_t, a)]$$

  requires $\widehat{r}$ to be consistent
- IS estimator

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{\widehat{b}(A_t|S_t)}R_t,$$

  requires $\widehat{b}$ to be consistent
- Doubly robust (DR) estimator combines both, and requires **either** $\widehat{r}$ **or** $\widehat{b}$ to be consistent (**"doubly-robustness" property**)

Doubly Robust Estimator (Cont'd)

- Consider the estimating function

$$\phi(S, A, R) = \sum_{a} \pi(a|S)\widehat{r}(S, a) + \frac{\pi(A|S)}{\widehat{b}(A|S)}[R - \widehat{r}(S, A)]$$

- First term on the RHS is the estimating function of the direct estimator
- Second term corresponds to the **augmentation term**
    - Zero mean when $\widehat{r} = r$
    - Debias the bias of the direct estimator
    - Offering additional robustness against model misspecification of $\widehat{r}$
- DR estimator given by $T^{-1}\sum_{t=0}^{T-1}\phi(S_t, A_t, R_t)$

Fact 1: Double Robustness

- The estimating function

$$\phi(S,A,R)=\sum_a \pi(a|S)\widehat{r}(S,a)+\frac{\pi(A|S)}{\widehat{b}(A|S)}[R-\widehat{r}(S,A)]$$

- In large sample size, DR estimator converges to $\mathbb{E}\phi(S,A,R)$
- When $\widehat{r}=r$, the augmentation term has zero mean. It follows that

$$\mathbb{E}\phi(S,A,R)=\sum_a \mathbb{E}[\pi(a|S)r(S,a)]=\text{target policy's value}$$

- When $\widehat{b}=b$, it has the same mean as the IS estimator

$$\begin{aligned}\mathbb{E}\phi(S,A,R)&=\mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)}R\right]+\mathbb{E}\left[\sum_a \pi(a|S)\widehat{r}(S,a)-\frac{\pi(A|S)}{b(A|S)}\widehat{r}(S,A)\right]\\&=\mathbb{E}\left[\frac{\pi(A|S)}{b(A|S)}R\right]=\text{target policy's value}\end{aligned}$$

Fact 2: Efficiency

- When $\widehat{\boldsymbol{b}} = \boldsymbol{b}$, the estimating function

$$\phi(S, A, R) = \sum_{a} \pi(a|S)\widehat{r}(S, a) + \frac{\pi(A|S)}{b(A|S)}[R - \widehat{r}(S, A)]$$

- The MSE of DR estimator is proportional to the variance of $\phi(S, A, R)$

$$\mathrm{Var}(\phi(S, A, R)) = \mathbb{E}[\mathrm{Var}(\phi(S, A, R)|S, A)] + \mathrm{Var}[\mathbb{E}(\phi(S, A, R)|S, A)]$$

- The first term on the RHS is independent of $\widehat{\boldsymbol{r}}$
- The second term is minimized when $\widehat{\boldsymbol{r}} = \boldsymbol{r}$
- A good working model for $\boldsymbol{r}$ improves the estimator's efficiency
- When $\widehat{\boldsymbol{r}} = \boldsymbol{r}$, the estimator achieves the **efficiency bound** [e.g., smallest MSE among a class of regular estimators; see Tsiatis, 2007]

Fact 3: Efficiency

- When $\widehat{\boldsymbol{b}}$ is estimated from data and the model is **correctly specified**, the estimator's MSE would be **generally smaller than** the one that uses the oracle behavior policy $\boldsymbol{b}$ [Tsiatis, 2007]
- Estimating $\widehat{\boldsymbol{b}}$ yields a more efficient estimator, even if we know the oracle $\boldsymbol{b}$
- **Multi-armed bandit** example without context information
  - **Objective**: evaluate $\mathbb{E}(\boldsymbol{R}|\boldsymbol{A} = \boldsymbol{a})$ for a given $\boldsymbol{a}$
  - IS estimator with **known** $\Pr(\boldsymbol{A} = \boldsymbol{a})$

$$\frac{\sum_{t=0}^{T-1} \mathbb{I}(\boldsymbol{A}_t = \boldsymbol{a})\boldsymbol{R}_t}{\boldsymbol{T}\Pr(\boldsymbol{A}_t = \boldsymbol{a})}$$

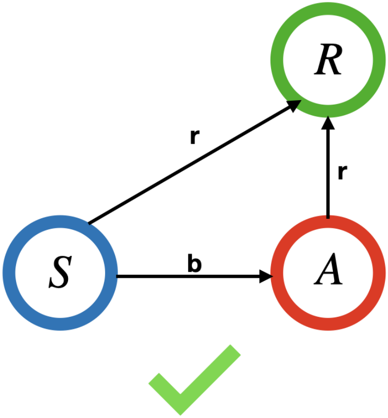
  - IS estimator with **estimated** $\Pr(\boldsymbol{A} = \boldsymbol{a})$ has a **smaller** asymptotic variance

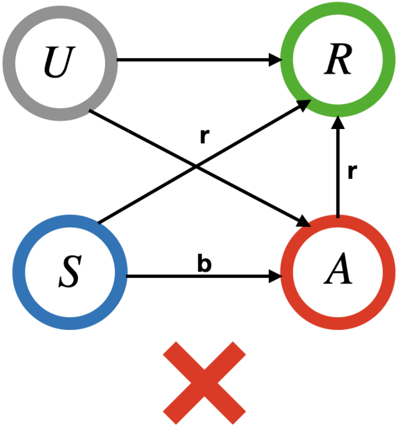
$$\frac{\sum_{t=0}^{T-1} \mathbb{I}(\boldsymbol{A}_t = \boldsymbol{a})\boldsymbol{R}_t}{\sum_{t=0}^{T-1} \mathbb{I}(\boldsymbol{A}_t = \boldsymbol{a})}$$

Fact 4: Asymptotic Normality

- The DR estimator converges at a parametric rate and is asymptotically normal even when both $\widehat{r}$ and $\widehat{b}$ converge **slower** than the parameter rate (i.e., root-n rate)
- This observation allows us to apply machine learning methods to estimate both nuisance functions, leading to the **double machine learning** estimator [Chernozhukov et al., 2017]
- Indeed, it only requires $\widehat{r}$ and $\widehat{b}$ to converge at a rate of $o_p(n^{-1/4})$, due to the double robustness property

Assumption: No Unmeasured Confounders





Lecture Outline

1. Off-Policy Evaluation (OPE) Introduction

2. OPE in Contextual Bandits

3. **OPE in Reinforcement Learning**

General OPE Problem

- **Objective**: Given an offline dataset $\{(S_{i,t}, A_{i,t}, R_{i,t}) : 1 \le i \le N, 0 \le t \le T\}$ generated by a behavior policy b, where i indexes the ith episode and t indexes the tth time point, we aim to evaluate the mean return under a target policy π

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] = \mathbb{E} V^{\pi}(S_0)$$

  When $\gamma = 1$, the task is assumed to be episodic
- We focus on the case where both π and b are **stationary** policies
- Challenge: **Distributional shift**
  - In the offline dataset, actions are generated according to b
  - The target policy π we wish to evaluate is different from b

Direct Estimator

- The target policy's value is given by $\mathbb{E} \boldsymbol{V}^{\pi}(\boldsymbol{S}_0)$, or equivalently,

$$\mathbb{E}[\sum_{\boldsymbol{a}} \pi(\boldsymbol{a}|\boldsymbol{S}_0)\boldsymbol{Q}^{\pi}(\boldsymbol{S}_0, \boldsymbol{a})]$$

- The expectation can be approximated via the **empirical initial state distribution**
- Q-learning is an **off-policy** algorithm. Can be applied to learn $\boldsymbol{Q}^{\pi}$ offline
- This yields the direct estimator

$$\frac{1}{\boldsymbol{N}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{N}} \sum_{\boldsymbol{a}} \pi(\boldsymbol{a}|\boldsymbol{S}_{\boldsymbol{i},0})\widehat{\boldsymbol{Q}}(\boldsymbol{S}_{\boldsymbol{i},0}, \boldsymbol{a})$$

- It remains to compute $\widehat{\boldsymbol{Q}}$

Fitted Q-Evaluation [Le et al., 2019]

- Bellman equation

$$\mathbb{E}\left[R_t + \gamma \pi(a|S_{t+1}) Q^{\pi}(S_{t+1}, a) | S_t, A_t\right] = Q^{\pi}(S_t, A_t)$$

- Both LHS and RHS involves Q^{π}
- Repeat the following procedure
  1. Compute $\widehat{Q}$ as the argmin of

$$\arg\min_{Q} \sum_{t} \left[R_{i,t} + \gamma \sum_{a} \pi(a|S_{i,t+1}) \widetilde{Q}(S_{i,t+1}, a) - Q(S_{i,t}, A_{i,t})\right]^2$$

  2. Set $\widetilde{Q} = \widehat{Q}$
- Designed for learning Q^{π}
- Do **not** require actions to follow the target policy

Other Direct Estimators

- Sieve-based estimator [Shi et al., 2020b]
  - Use linear sieves to parametrize $\boldsymbol{Q^{\pi}}$
  - Estimate regression coefficients by solving the Bellmen equation
- Kernel-based estimator [Liao et al., 2021]
  - Use RHKSs to parametrize $\boldsymbol{Q^{\pi}}$
  - Estimate parameters by solving a coupled optimization [Farahmand et al., 2016]
- Limiting distributions of value estimators are derived in the two papers

Stepwise IS Estimator [Zhang et al., 2013]

- Consider episodic task where T is the termination time
- Importance sampling ratio needs to be employed

$$
\begin{aligned}
\mathbb{E}^{\pi} R_0 &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} R_0\right] \\
\mathbb{E}^{\pi} R_1 &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \frac{\pi(A_1|S_1)}{b(A_1|S_1)} R_1\right] \\
&\vdots \\
\mathbb{E}^{\pi} R_t &= \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \cdots \frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_t\right]
\end{aligned}
$$

Stepwise IS Estimator (Cont'd)

- According to this logic, the target policy's value can be represented by

$$\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left\{\prod_{j=0}^{t} \frac{\pi(A_j|S_j)}{b(A_j|S_j)}\right\} R_t\right]$$

- This yields the stepwise IS estimator

$$\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T} \gamma^t \left\{\prod_{j=0}^{t} \frac{\pi(A_{i,j}|S_{i,j})}{\widehat{b}(A_{i,j}|S_{i,j})}\right\} R_{i,t}\right]$$

for a given estimator $\widehat{b}$ computed using supervised learning algorithms

Limitation

- Stepwise IS suffers from a **large variance**
- In particular, the IS ratio at time t is the product of individual ratios from the **initial** time to time t

$$\prod_{j=0}^{t} \frac{\pi(A_j|S_j)}{b(A_j|S_j)}$$

- Variance of the ratio grows **exponentially** with respect to t, referred to as the **curse of horizon** [Liu et al., 2018]
- Extension: **Doubly-robust** estimator by [Jiang and Li, 2016]

Pros & Cons of Direct v.s. Stepwise IS

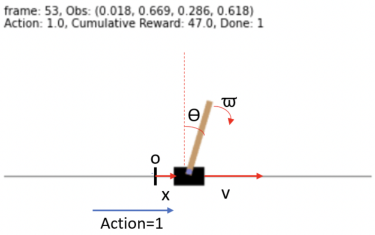
- Bias/Variance Trade-Off
- When $\boldsymbol{b}$ is known, stepwise IS is an **unbiased** estimator since

$$\mathbb{E}^{\pi} R_t = \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)} \cdots \frac{\pi(A_t|S_t)}{b(A_t|S_t)} R_t\right]$$

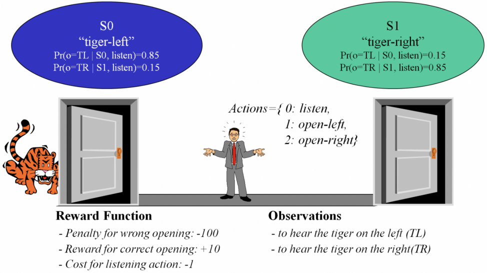
- Direct estimator has **some bias**, since Q^{π} needs to be estimated from data
- Stepwise IS suffers from **curse of horizon** and a **large variance**
- Direct estimator has a much lower variance

Pros & Cons of Direct v.s. Stepwise IS (Cont'd)

- Direct estimator exploits **Markov** & **stationary** properties
- Relies on the **Bellman equation**
- More **efficient** in MDP environments



- SIS does **not** exploit these properties
- More **flexible** in non-MDP environments (e.g., POMDP)



Marginalized IS Estimator

- As we have discussed, stepwise IS suffers from **curse of horizon**
- Curse of horizon is **unavoidable** in general **Non-Markov decision processes** (e.g., POMDP)
- Under some additional model assumptions (e.g., Markovianity & time-homogeneity), it is possible to break the curse of horizon using **marginalized IS** estimator
- Stepwise IS does **not** exploit these properties

Marginalized IS Estimator (Cont'd)

- Stepwise IS uses the **cumulative** IS ratio

$$\mathbb{E}^{\pi} R_t = \mathbb{E}^{b}\left[\frac{\pi(A_0|S_0)}{b(A_0|S_0)}\cdots\frac{\pi(A_t|S_t)}{b(A_t|S_t)}R_t\right]$$

- Under Markovianity (TMDP), marginalized IS uses the **marginalized** IS ratio

$$\mathbb{E}^{\pi} R_t = \mathbb{E}^{b}\left[\frac{p_t^{\pi}(S_t, A_t)}{p_t^{b}(S_t, A_t)}R_t\right] \tag{1}$$

  where p_t^{π} and p_t^{b} are the marginal density functions of (S_t, A_t) under π and b
- The resulting marginalized IS estimator can be derived from (1)

Marginalized IS Estimator

- Under Markovianity and time-homogeneity (MDP),

$$\mathbb{E}V^{\pi}(S_0) = \mathbb{E}^{b}\left[\frac{\sum_{t=0}^{\infty}\gamma^t p_t^{\pi}(S,A)}{p_{\infty}(S,A)}R\right] \tag{2}$$

  where p_{∞} denotes the limiting state-action distribution under b and the numerator corresponds to the γ-discounted state-action visitation probability
- The resulting marginalized IS estimator can be derived from (2)
- Marginal IS ratio can be estimated via **minimax learning** [Uehara et al., 2019]
- Closed-form expression is available when using **linear sieves**
- Coupled optimization can also be employed when using **RKHSs** [Liao et al., 2020]
- Alternatively, we can use **RKHSs** to parametrize the discriminator class, use **neural networks** to parametrize the ratio and apply SGD for parameter estimation

Double RL [Kallus and Uehara, 2019]

- Double RL extends DR in **contextual bandits** to the general RL problem
- Similar to DR, the estimator can be represented as

$$\text{Direct Estimator} + \text{Augmentation Term}$$

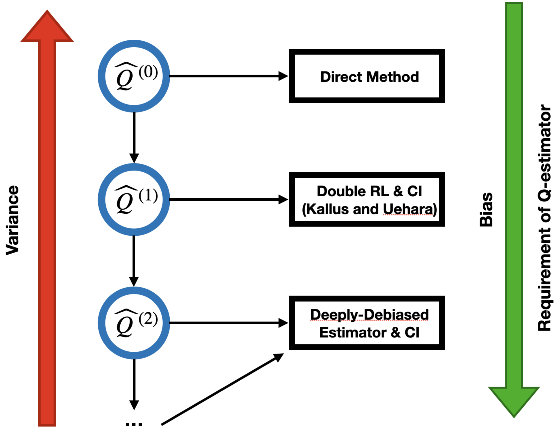
- **Augmentation** term is to **debias** the bias of direct estimator and offer protection against model misspecification of $\boldsymbol{Q^{\pi}}$; it relies on the marginalized IS ratio
- Similar to DR, the estimator is **doubly-robust**, e.g., consistent when either $\boldsymbol{Q^{\pi}}$ or the marginalized IS ratio is correct
- Similar to DR, the estimator achieves the **efficiency bound** in MDPs

Fact 5: Efficiency

- Direct estimators (based on linear sieves or RKHSs) also achieve the **efficiency bound** in MDPs [Liao et al., 2021, Shi et al., 2022a]
- Marginalized IS estimators (based on linear sieves) also achieve the **efficiency bound** in MDPs
- When using linear sieves,

$$\text{direct estimator} = \text{marginalized IS estimator} = \text{double RL estimator}$$

Deeply-Debiased OPE [Shi et al., 2021b]



- Constructed based on high-order influence function [Robins et al., 2008, 2017]
- Ensures bias decays much faster than standard deviation
- Allows to provide valid **uncertainty quantification** (e.g., confidence interval)

Other Topics

- Evaluation of the expected return under optimal policy
  - Inference is challenging in **nonregular** settings where the optimal policy is not unique
  - m-out-of-n bootstrap [Chakraborty et al., 2013]
  - Martingale-based method [Luedtke and Van Der Laan, 2016, Shi et al., 2020b]
  - Subagging-based method [Shi et al., 2020a]
- Confounded OPE
  - Confounded POMDPs [Tennenholtz et al., 2020, Bennett and Kallus, 2021, Shi et al., 2021a]
  - Confounded MDPs [Zhang and Bareinboim, 2016, Wang et al., 2021, Fu et al., 2022, Shi et al., 2022b]

References I

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

Bibhas Chakraborty, Eric B Laber, and Yingqi Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3): 714–723, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

References II

Zuyue Fu, Zhengling Qi, Zhaoran Wang, Zhuoran Yang, Yanxun Xu, and Michael R Kosorok. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*, 2019.

Hoang M Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*, 2020.

References III

Peng Liao, Predrag Klasnja, and Susan Murphy. Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association*, 116(533):382–391, 2021.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

References IV

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, Aad van der Vaart, et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

Chengchun Shi, Wenbin Lu, and Rui Song. Breaking the curse of nonregularity with subagging: inference of the mean outcome under optimal treatment regimes. *Journal of Machine Learning Research*, 21, 2020a.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020b.

Chengchun Shi, Masatoshi Uehara, and Nan Jiang. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021a.

References V

Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR, 2021b.

Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, (just-accepted):1–29, 2022a.

Chengchun Shi, Jin Zhu, Ye Shen, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *arXiv preprint arXiv:2202.10589*, 2022b.

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283, 2020.

References VI

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

References VII

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.