Estimating Long-term Rewards by Off-policy Reinforcement Learning

Lihong Li
Amazon

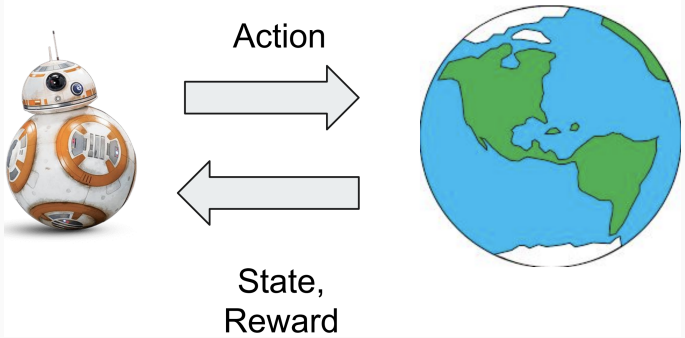
November 10th, 2021

OSU AI Seminar

Thanks to Andrew Bennett, Yinlam Chow, Bo Dai, Nathan Kallus, Qiang Liu, Ali Mousavi, Ofir Nachum, Dale Schuurmans, Csaba Szepesvári, and many other collaborators

RL: A General Learning Paradigm

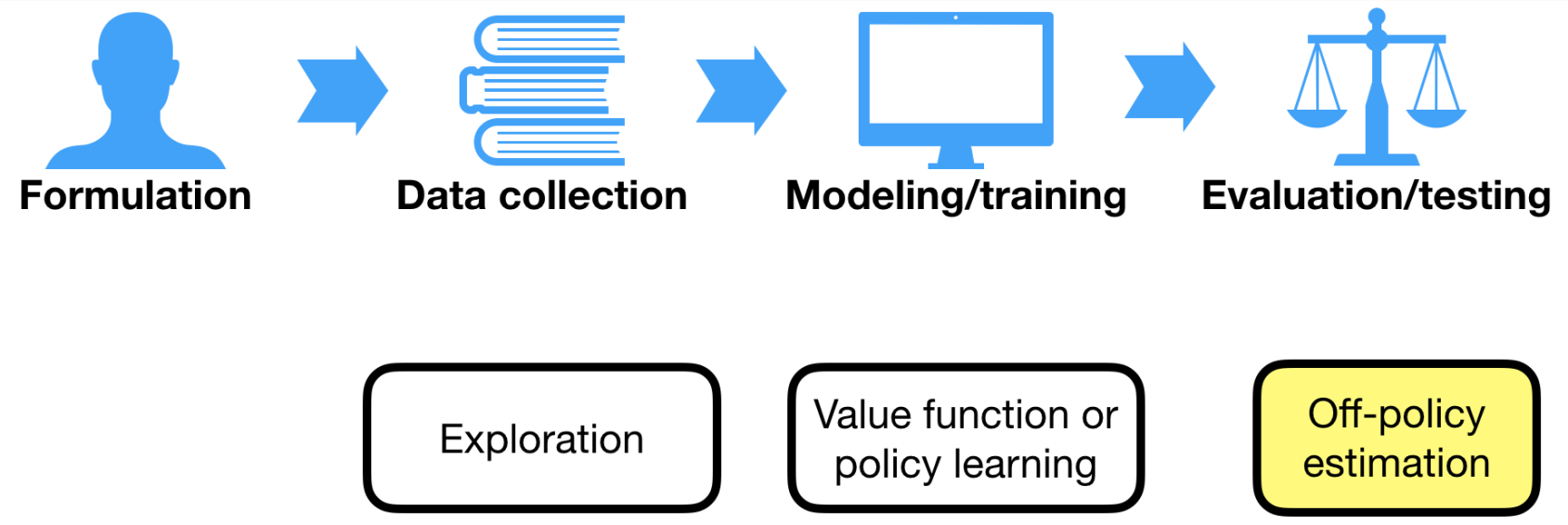
- Games
- Robotics and control
- Web: advertising, recommendation, search
- Inventory management
- Scheduling and planning
- Conversational systems
- ...



A Simplified View of RL

customer problem

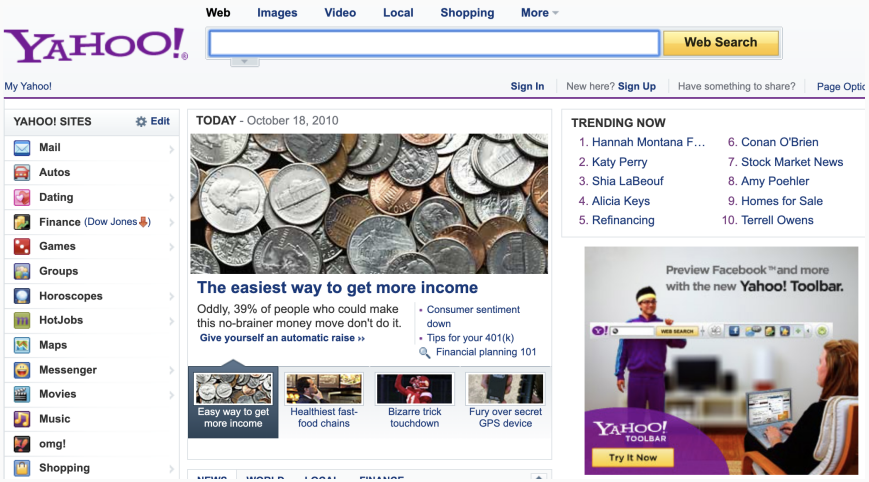
$\Downarrow$



Informally, off-policy estimation (OPE) asks the question:

How good is a given policy without actually running it?

OPE Example: Personalized News Recommendation



(www.yahoo.com on 2010/10/18)

Naturally an RL problem:

- state: user information
- action: news article
- reward: $\{0, 1\}$ (click or not)
- policy: state $\rightarrow$ action

A typical <u>evaluation dilemma</u> in practice:

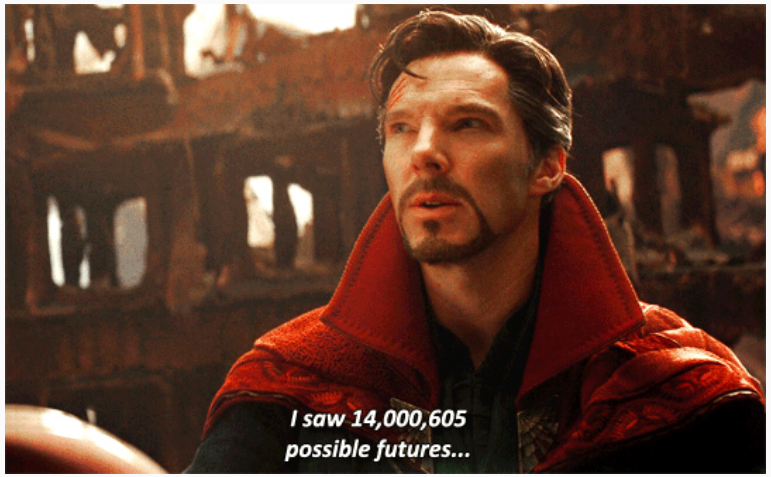
- policy deployment requires performance evaluation
- performance evaluation requires policy deployment

OPE: Can we evaluate performance without policy deployment?

Fundamentally a counterfactual estimation problem.

OPE with Simulators

- Games: (almost) perfect simulation
- Robotics: physical dynamics
- Recommendation: user click model
- Retail: consumer behavior model
- Auto. driving: all traffic conditions
- Dialogue: human conversation model
- Healthcare: patient's medical model
- ...



Simulators can be a very useful tool.

But reliable simulators are often hard to build.

This talk: complementary approaches **without** a simulator.

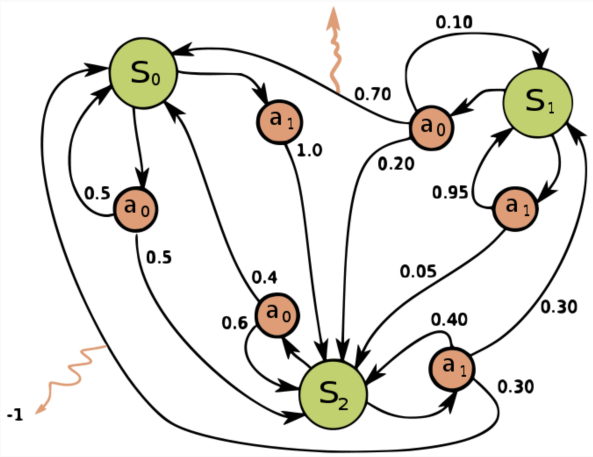
Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

Markov Decision Process

MDP $M = \langle \mathcal{S}, \mathcal{A}, T, R, \mu_0, \gamma \rangle$

- $\mathcal{S}$: finite set of states
- $\mathcal{A}$: finite set of actions
- $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: transition kernel
- $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$: reward function
- $\mu_0 \in \Delta(\mathcal{S})$: initial state distribution
- $\gamma \in [0, 1)$: discount factor



(from Wikipedia)

See papers for treatments of infinite state-actions, and of the $\gamma = 1$ case.

Definitions

- Trajectory $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, \ldots)$, with $s_0 \sim \mu_0$ controlled by policy π: $\forall t,\ a_t \sim \pi(\cdot|s_t),\ s_{t+1} \sim T(\cdot|s_t, a_t)$
- Return: $U(\tau) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t r_t$
- Policy value: $v(\pi) := \mathbb{E}_{\tau \sim \pi}[U(\tau)]$ $\Longleftarrow$ **quantity to estimate**

Estimating $v(\pi)$

On-policy (Monte Carlo) estimation

- Run π directly to generate trajectories
- Estimate $v(\pi)$ (by averaging rewards over trajectories)

E.g., randomized clinical trials, A/B testing, ...

Statistically sound, and straightforward

Often costly, risky, or even impossible to run π

Examples: healthcare, autonomous driving, ...

Off-policy estimation

- Assume historical data $\bar{\mathcal{D}}$ generated by behavior policy π_0
- Estimate $v(\pi)$ from $\bar{\mathcal{D}}$

Cheap and safe

Biased: simple average over $\bar{\mathcal{D}}$ estimates $v(\pi_0)$, not $v(\pi)$

Connection to Causal Effect Estimation

A Special Case

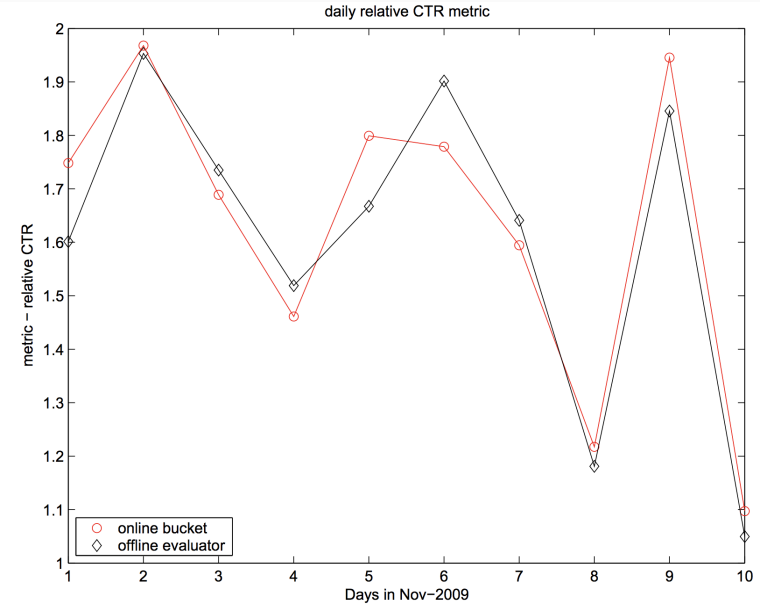
- $\mathcal{S}$ is set of patients
- $\mathcal{A} = \{\texttt{NewDrug}, \texttt{Placebo}\}$
- Reward is $\{0, 1\}$ (cured or not)
- Discount $\gamma = 0$
- $\pi_{\texttt{treatment}}(s) \equiv \texttt{NewDrug}$,
  $\pi_{\texttt{control}}(s) \equiv \texttt{Placebo}$
- Want to estimate $v(\pi_{\texttt{treatment}}) - v(\pi_{\texttt{control}})$

Same as the average treatment effect.

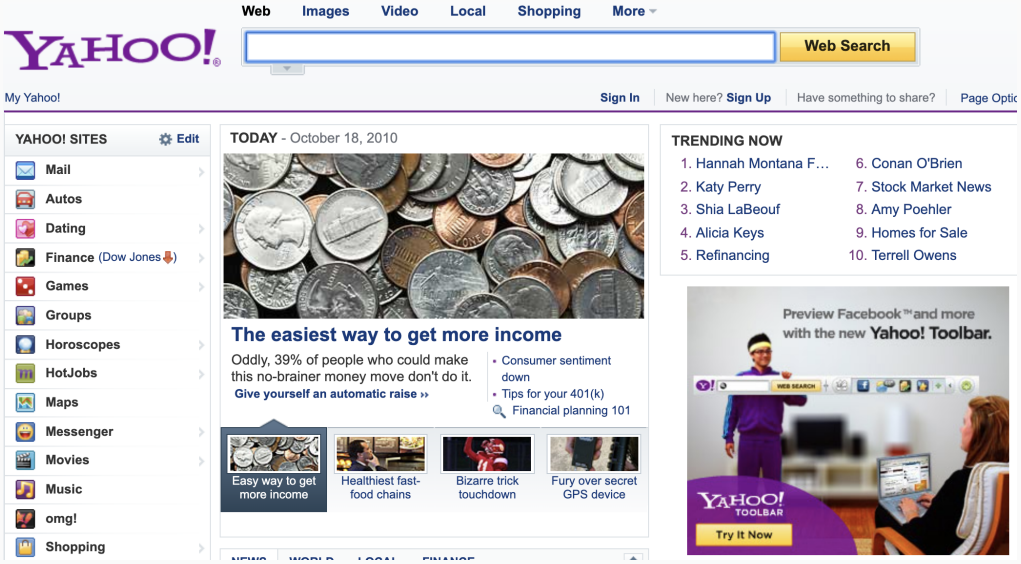
Also closely related to covariate shift in machine learning.

Personalized News Recommendation

- $\mathcal{S}$ is all users
- $\mathcal{A}$ is news articles
- Reward is $\{0, 1\}$ (click or not)
- $\gamma = 0$ (i.e., contextual bandits)
- $v(\pi)$ is click-through rate



[LCLW'11]



Survey on more applications on the Web [HLR'16]

Family Functioning

Fast Track prevention program [Murphy'03]

- Goal: prevent/reduce conduct disorder of children
- $\mathcal{S}$: child's behavioral condition at the end of semester
- $\mathcal{A}$: number of home visits scheduled in the next semester
- Reward: academic achievement in 10 semesters
- $\gamma = 1$

Each step $t \in \{0, 1, \ldots, 9\}$ corresponds to 1 semester.

Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

Precise Formulation

Generation of data

$$\begin{aligned} \bar{\mathcal{D}} &= \left\{\tau^{(i)}\right\}_{1\le i\le m} \\ \tau^{(i)} &= \left(s_0^{(i)}, a_0^{(i)}, r_0^{(i)}, \ldots, s_{H-1}^{(i)}, a_{H-1}^{(i)}, r_{H-1}^{(i)}\right), \end{aligned}$$

for some large enough horizon H and behavior policy π_0, where $a_t \sim \pi_0(\cdot|s_t)$, $r_t = R(s_t, a_t)$, $s_{t+1} \sim T(\cdot|s_t, a_t)$.

OPE: Off-policy estimation

Estimate $\hat{v}(\bar{\mathcal{D}}, \pi) \approx v(\pi)$ **without** knowing $\{T, R\}$.

When π_0 is unknown, "behavior-agnostic OPE".

Bellman Equation

Value functions

Generalizing policy value:

$$Q_\pi(s,a) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r_t | s_0 = s, a_0 = a]$$

So $v(\pi) = (1-\gamma)\mathbb{E}_{s_0\sim\mu_0}[Q_\pi(s_0,\pi(a_0))]$.

Bellman equation

$$Q_\pi(s,a) \;=\; R(s,a) + \gamma\mathbb{E}_{s'\sim T(\cdot|s,a),a'\sim\pi(\cdot|s)}[Q_\pi(s',a')]$$

Compute Q_π from known (T,R): dynamic programming, ...

Regression Estimator

A plug-in approach

- Use $\bar{\mathcal{D}}$ to estimate (T, R) of MDP by $(\hat{T}, \hat{R})$
- Compute $\hat{Q}_\pi$ from $(\hat{T}, \hat{R})$, and estimate $v(\pi)$ by

$$\begin{aligned}\hat{v}_{\text{Reg}} &= (1-\gamma)\hat{\mathbb{E}}_{s_0}[\hat{Q}_\pi(s_0, \pi(s_0))] \\ &= \frac{1-\gamma}{m} \sum_{\tau \in \bar{\mathcal{D}}} \hat{Q}_\pi(s_0, \pi(s_0))\end{aligned}$$

Pros: simple and low variance.

Cons: uncontrolled bias due to model misspecification of $(\hat{T}, \hat{R})$.

Inverse Propensity Score (IPS)

Key idea: change of measure

$$
\begin{aligned}
v(\pi) &= \mathbb{E}_{\tau\sim\pi}[U(\tau)] \\
&= \mathbb{E}_{\tau\sim\pi_0}\left[\frac{\Pr_\pi(\tau)}{\Pr_{\pi_0}(\tau)}U(\tau)\right] \\
&= \mathbb{E}_{\tau\sim\pi_0}\left[\frac{\prod_{t=0}^{H-1} T(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)}{\prod_{t=0}^{H-1} T(s_t|s_{t-1},a_{t-1})\pi_0(a_t|s_t)}U(\tau)\right] \\
&\approx \frac{1}{|\bar{\mathcal{D}}|}\sum_{\tau\in\bar{\mathcal{D}}}\left[\underbrace{\prod_{t=0}^{H-1}\frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}}_{\text{“propensity score”}} U(\tau)\right] =: \hat{v}_{\text{IPS}}\,.
\end{aligned}
$$

Pros: unbiased.

Cons: high variance (more discussion soon).

Many Variants of IPS/DR

- Per-decision IPS [Precup+'01]
- Self-normalized IPS [Precup+'01, Swaminathan+'15]
- Doubly robust estimators [J**L**'16, Thomas+'16]
- More robust doubly robust estimators [Farajtabar+'18]
- Use of temporal abstraction to reduce variance [Guo+'17]
- ...

Unfortunately, not widely used in practice unless $\gamma \ll 1$.

Curse of Horizon

- For $\tau = (s_0, a_0, r_0, \ldots, s_{H-1}, a_{H-1}, r_{H-1})$,

$$\frac{\Pr_\pi(\tau)}{\Pr_{\pi_0}(\tau)} = \prod_{t=0}^{H-1} \frac{\pi(a_t|s_t)}{\pi_0(a_t|s_t)}$$

  whose variance easily blows up (exponentially in horizon H).
- Exponential blow-up is unavoidable in *minimax* rate [LMS'15]
- Similar for doubly robust estimator [JL'16]
- Can truncate ratios when they grow too large, at the cost of introducing large bias (e.g. ReTrace [Munos+'16])

Can we break the Curse of Horizon (at least in "benign" cases)?
Many recent developments [LLTZ'18, NCDL'19, DNCLSS'20, ...].

Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

Behavior-agnostic OPE

Now consider a more general setting where data consists of n transition tuples: $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{1 \le i \le n}$

Assumptions:

- $(s_i, a_i) \sim d_{\mathcal{D}}$ (unknown behavior distribution)
- $\mathbb{E}[r_i] = R(s_i, a_i)$
- $s_i' \sim T(s_i, a_i)$

Otherwise, no assumption on π_0 (other than regularity conditions). Data do not even have to be in the form of trajectories.

Linear Program Characterizations of Policy Value

Two equivalent LPs for the policy value [Puterman'94]:

Primal LP

$$\begin{aligned} v(\pi) = \quad & \min_{Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad (1-\gamma)\,\mathbb{E}_{(s_0,a_0)\sim\mu_0\pi}[Q(s_0,a_0)] \\ & \text{s.t.} \qquad Q \geq R + \gamma\mathcal{P}_\pi Q \end{aligned}$$

Dual LP

$$\begin{aligned} v(\pi) = \quad & \max_{d\in\Delta(\mathcal{S}\times\mathcal{A})} \quad \mathbb{E}_{(s,a)\sim d}[R(s,a)] \\ & \text{s.t.} \qquad d = (1-\gamma)\mu_0\pi + \gamma\mathcal{P}_\pi^* d \end{aligned}$$

where

$$\mathcal{P}_\pi Q(s,a) := \sum_{s',a'} T(s'|s,a)\pi(a'|s')Q(s',a')$$

$$\mathcal{P}_\pi^* d(s,a) := \sum_{\bar{s},\bar{a}} \pi(a|s)T(s|\bar{s},\bar{a})d(\bar{s},\bar{a})$$

OPE based on Dual LP

$$v(\pi) = \max_{d \in \Delta(\mathcal{S} \times \mathcal{A})} \mathbb{E}_{(s,a) \sim d}[R(s,a)]$$
$$\text{s.t.} \quad d = (1-\gamma)\mu_0 \pi + \gamma \mathcal{P}_\pi^* d \qquad (*)$$

Observations

- There is a unique solution $d = d_\pi$ that satisfies (*).
  Known as stationary distribution, steady-state distribution, ...
- A dual approach to obtaining policy value:

$$v(\pi) = \mathbb{E}_{d_\pi}[R(s,a)] = \mathbb{E}_{d_\mathcal{D}}[\zeta_\pi(s,a) R(s,a)]$$

  where $d_\mathcal{D}$ is the data distribution,
  and $\zeta_\pi(s,a) = \frac{d_\pi(s,a)}{d_\mathcal{D}(s,a)}$ is Radon-Nikodym derivative (RND)
  (under absolute continuity condition $d_\pi \ll d_\mathcal{D}$)
- OPE is now reduced to estimating $\zeta(s,a)$

DualDICE Approach to OPE

Observation

ζ_π is the solution to a large linear system:

$$\zeta d_{\mathcal{D}} = (1-\gamma)\mu_0 \pi + \gamma \mathcal{P}_\pi^* \zeta d_{\mathcal{D}}$$

- The dimension ($|\mathcal{S} \times \mathcal{A}|$) is large or even ∞. We may use models $\hat{\zeta}(s, a; \theta)$ with parameter θ (e.g., NNs)
- We need to approximate $d_{\mathcal{D}}$ and $\mathcal{P}_\pi^*$ with empirical data (and proper regularization is necessary)

Solution

DualDICE (Dual DIstribution Correction Estimation) formulates it into an optimization problem to solve for ζ_π [NCD**L**'19]:

$$\text{Solve for} \quad \hat{\zeta} \leftarrow \min \text{Discrepancy}\left((1-\gamma)\mu_0 \pi + \gamma \mathcal{P}_\pi^* \zeta d_{\mathcal{D}} - \zeta d_{\mathcal{D}}\right)$$

More on DualDICE

Convergence [NCDL'19]

Under proper assumptions, the mean squared error

$$\mathbb{E}\left[\hat{v}_{\text{DualDICE}}(\pi) - v(\pi)\right]^2 = \tilde{O}\left(\varepsilon_{\text{approx}} + \varepsilon_{\text{opt}} + \frac{1}{\sqrt{n}}\right)$$

Variations and Extensions

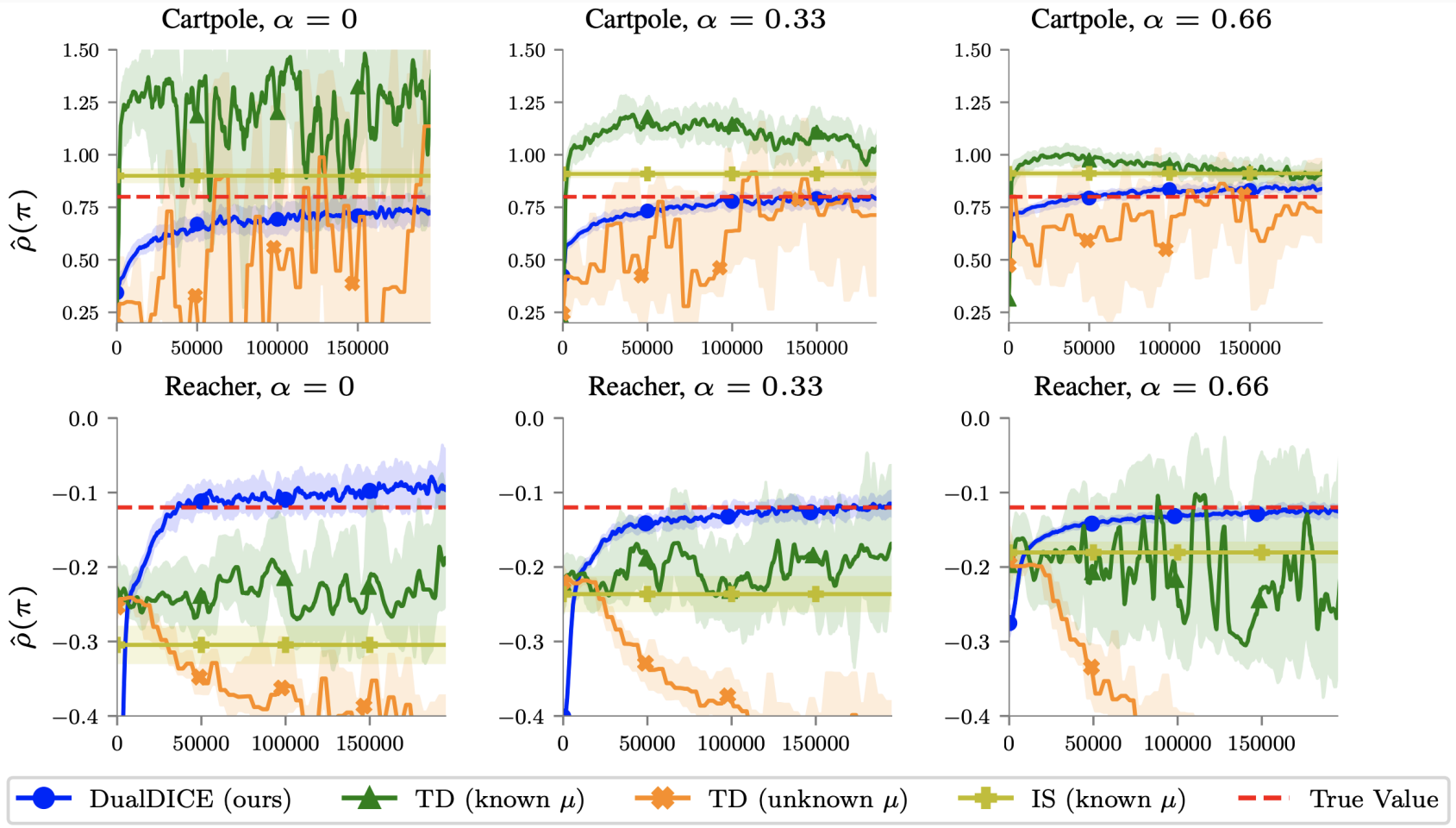
- Extension to $\gamma = 1$ by GenDICE [ZDLS'20]
- Doubly robustness [ZFLZL'20,Uehara+'20]
- A unification [YNDLS'20] (DualDICE, GenDICE, LSTDQ [Lagoudakis&Parr'03], MQL/MWL [Uehara+'20], ...)
- Varitional power iteration for easier optimization [PDLS'20]
- Consistent policy gradient estimation (for policy optimization) from off-policy data [NDKCLS'19]

Numerical Experiments

MuJoCo: Multi-Joint dynamics with Contact.

(from mujoco.org)

DualDICE Numerical Experiments



- Mujoco: simulated control
- α controls the degree of "off-policy-ness" ($d_\pi = d_\mathcal{D}$ if $\alpha = 1$)

Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

Uncertainty Estimation

Point estimate alone (like DualDICE) has limited use.
Variance matters in most applications!

Popular methods

- Normal approximation based on CLT [Bottou+'13, LCKG'15]
- Concentration inequalities [Thomas+'15]
- Bootstrap [Thomas+'15; Hanna+'17]

CoinDICE

Our approach: CoinDICE [DNCLSS'20]

- Based on empirical likelihood method
- Computationally tractable
- Avoids the curse of horizon

Function space embedding of constraints:

$$\cancel{v(\pi)}\ \tilde{v}(\pi) = \max_{d \in \Delta(\mathcal{S}\times\mathcal{A})} \mathbb{E}_{(s,a)\sim d}[R(s,a)]$$

$$\text{s.t.} \quad \cancel{d = (1-\gamma)\mu_0\pi + \gamma\mathcal{P}_\pi^* d}$$

$$\langle \phi,\ (1-\gamma)\mu_0\pi + (\gamma\mathcal{P}_\pi^* - I)d \rangle = 0$$

for some given feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^p$.

Constraints become generalized estimating equations (GEE).

CoinDICE: Empirical likelihood

Basic ideas:

- Previously, for point estimates, we approximated $d_{\mathcal{D}}$ and $\mathcal{P}_{\pi}^{*}$ with empirical data $\mathcal{D}$ to solve the GEE (as in DualDICE)
- Now introduce w: w_i is the weight of the ith transition in $\mathcal{D}$
- Solve for ζ using the w-weighted GEE, for w within a neighborhood of uniform weighting
- The range of solved ζ induces an (approximate) confidence interval for $\tilde{v}(\pi)$

CoinDICE: Statistical Coverage

Asymptotic coverage
Under proper assumptions:

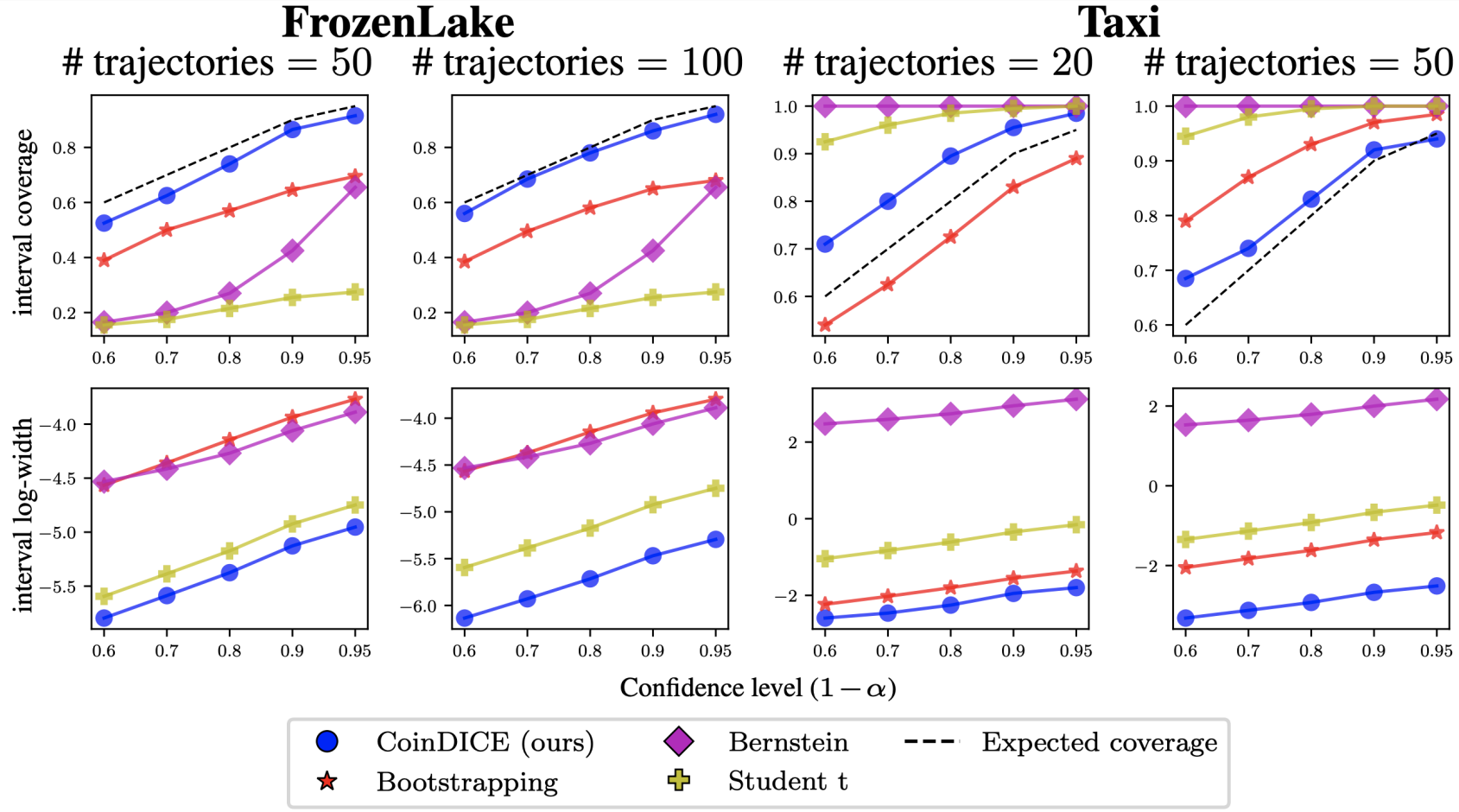
$$\lim_{n\to\infty} \Pr\left(\tilde{v}(\pi) \in C_{n,\xi}\right) = \Pr\left(\chi^2_{(1)} \leq \xi\right)$$

Non-asymptotic coverage
$C_{n,\xi}$ is a valid confidence interval, up to error of order $O(n^{-1})$.

Word of caution
The CI considers statistical error, but not approximation error.

It applies to $\tilde{v}$, not $v(\pi)$.

CoinDICE Numerical Experiments

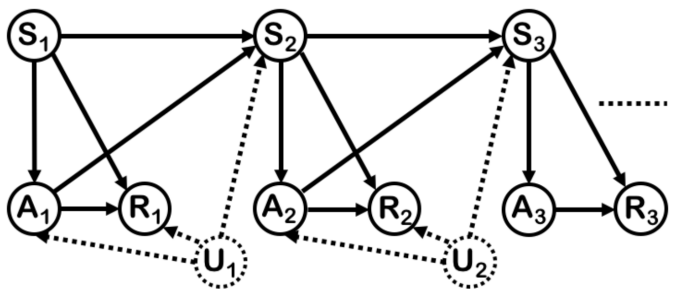


- Two standard benchmark problems
- α is the intended confidence interval coverage

Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

MDPs with Unmeasured Confounding



- Described by tuple $\langle \mathcal{S}, \mathcal{A}, ..., \mathcal{U} \rangle$
- $(u_1, u_2, u_3, \cdots)$ are IID
- Same as MDP except for $\Pr(s'|s, a, u)$, $R(s, a, u)$, $\pi_0(a|s, u)$ and $\pi(a|s, u)$
- MDPs $\subset$ MDPUCs $\subset$ partially observable MDPs
- Simplified example: insulin injection for type-1 diabetes patients
  - $\mathcal{S}$: blood glucose levels
  - $\mathcal{U}$: exogenous unmeasured events (e.g., food intake, exercise)

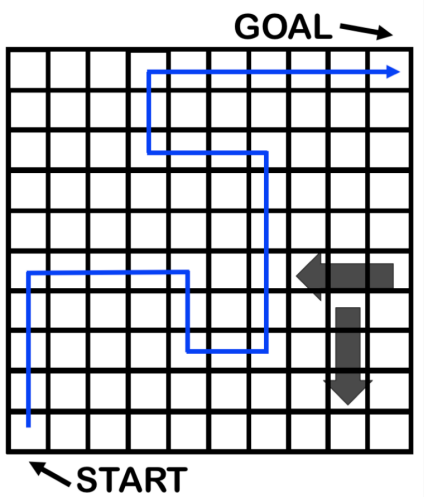
OPE in MDPUC

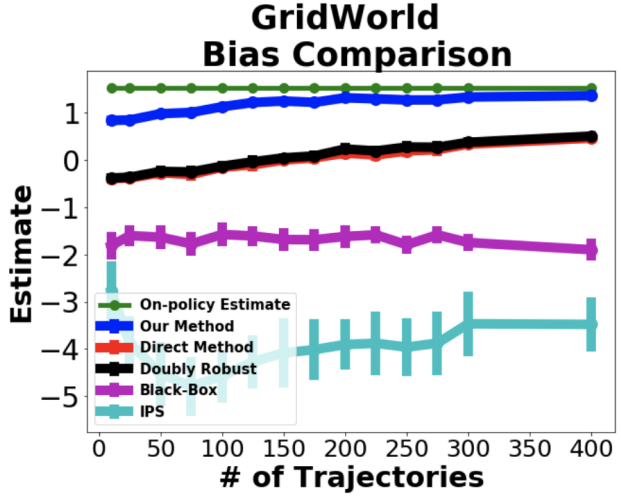
- Same form of estimator:

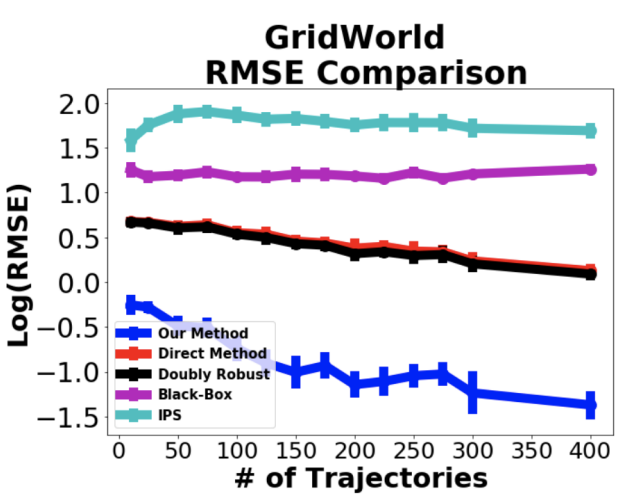
$$\hat{v}_{\text{MDPUC}} = \frac{1}{n} \sum_{i=1}^{n} \zeta_i r_i$$

  with ζ_i is the importance ratio as before (u gets canceled thanks to IID assumption)
- Previous method can be extended (under certain technical assumptions such as ergodicity), to achieve $1/n^{-1/2}$ convergence rate.
- See paper for details [BK**L**M'20]

Experiments: Windy Gridworld







- Binary confounders that affect behavior policy, reward and transition probabilities
- Our method (blue) consistently estimates the true value, but not other baselines
- Sensitivity analysis and more results in the paper [BKLM'20]

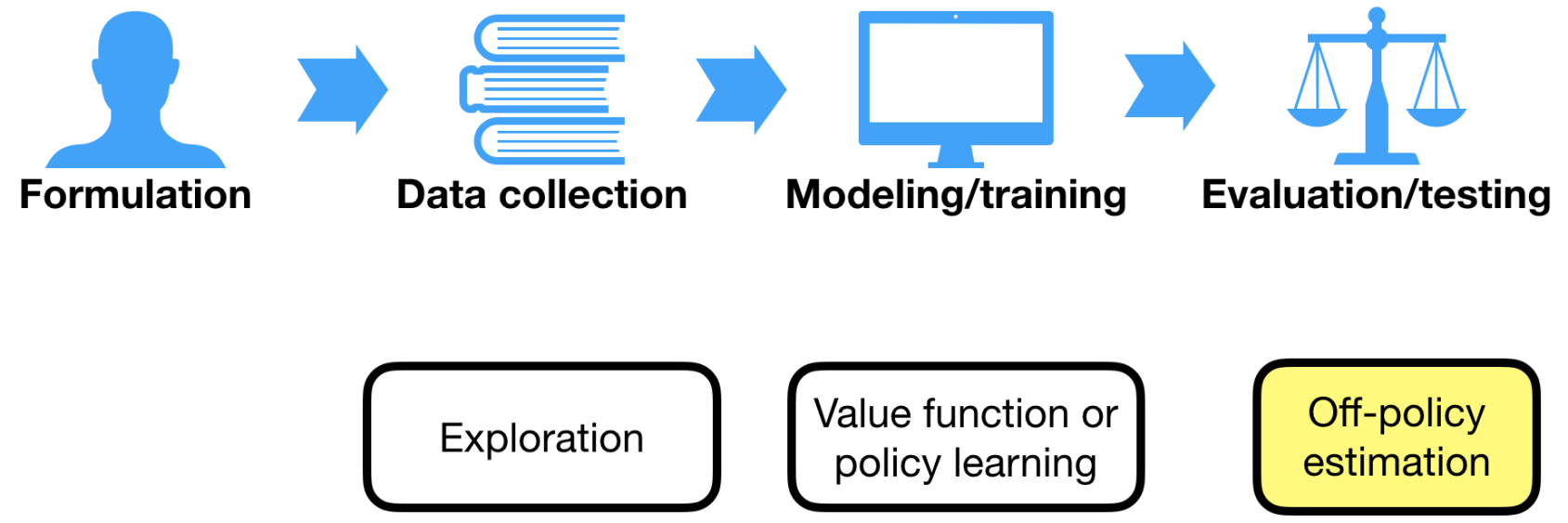
Outline

- The OPE problem
- Basic techniques
- OPE for long-term rewards
  - DualDICE: efficient point estimation
  - CoinDICE: confidence interval estimation
  - Dealing with unmeasured confounding
- Conclusions

A Simplified View of RL

customer problem

$\Downarrow$



Conclusions

Recap

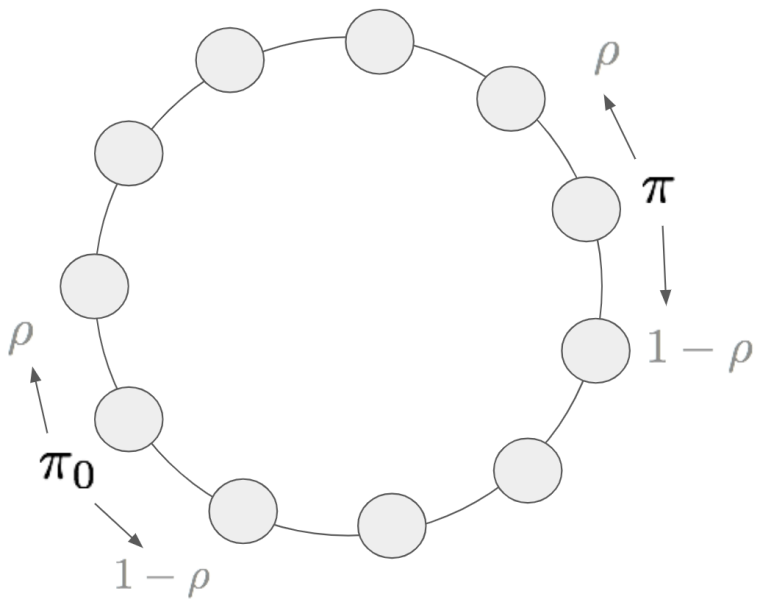
- Off-policy estimation is a critical bottleneck in "real-life" RL
- Many successes, mostly in short-horizon problems
- We break the curse of horizon based on duality structure in RL
- Extensions to confidence interval, and confounded cases

Future work

- OPE in more general confounded cases
- OPE in multiagent systems
- Applications in real-world problems

THANK YOU!

Is Importance-Weighting Trajectories Necessary?



See [LLTZ'18] for details.

"Circle" MDP

- $n = |\mathcal{S}|$ states arranged on circle
- Two actions:
  clockwise, counterclockwise
- Deterministic transitions

As $H \to \infty$:

- IPS/DR variance goes to ∞ (as the trajectory-wise importance weight's variance goes to ∞)
- But both policies visit every state equally often (as the state-wise importance weight is simply 1!)

Selected References Mentioned in the Slides

(Click on the title for the paper link)

- [B+'13] Counterfactual reasoning and learning systems: the example of computational advertising
- [BKLM'20] Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders
- [DLL'11] Doubly robust policy evaluation and learning
- [DNCLSS'20] CoinDICE: Off-policy confidence interval estimation
- [FTZL'21] Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds
- [HLR'16] Online evaluation for information retrieval, Chapter 4
- [JL'11] Doubly robust off-policy value evaluation for reinforcement learning
- [LCLW'11] Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms
- [LLTZ'18] Breaking the curse of horizon: Infinite-horizon off-policy estimation
- [NCDL'19] DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections