CoinDICE: Off-policy Confidence Interval Estimation

Bo Dai

Google Research, Brain Team

joint work with Ofir Nachum, Yinlam Chow, Lihong Li,
Csaba Szepesvári and Dale Schuurmans

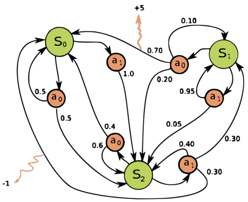
Overview

Table of Content

Markov Decision Processes (MDPs)

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

- $\mathcal{S}$: (possible infinite) set of states
- $\mathcal{A}$: (possible infinite) set of actions
- $T(s'|s, a)$: transition probabilities
- $R(s, a)$: immediate reward
- $\gamma \in (0, 1]$: discounted factor
- $\mu_0 \in \mathcal{P}(\mathcal{S})$: initial state distribution



Terminology

- Policy: $\pi(\cdot|s) : \mathcal{S} \to \mathcal{P}(\mathcal{A})$
- Trajectoy: $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$
- Return: $U(\tau) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t r_i$
- Value of policy: $v(\pi) = \mathbb{E}[U(\tau)]$

Off-policy Policy Evaluation

- **Historic experiences**:

$$
\begin{aligned}
\mathcal{D} &= \{x^i\}_{i=1}^m \\
x^i &= (s_0, a_0, s, a, r, s', a'),
\end{aligned}
$$

  with $(s_0, a_0) \sim \mu_0 \pi$, $(s, a, r, s') \sim d^{\mathcal{D}}$, and $a' \sim \pi(\cdot|s')$ where $d^{\mathcal{D}}$ is an unkown distirbution induced by some policies.

- **Goal**: Estimate $\widehat{v}(\mathcal{D}, \pi) \approx v(\pi) = \mathbb{E}_{\tau \sim \pi}\left[U(\tau)\right]$ **without** knowing T and R.
- If the behavior policies inducing $d^{\mathcal{D}}$ is also unknown, the task is called **bahavior-agnostic OPE**.

Table of Content

Linear Programming for Policy Value

Primal

$$\min_{Q:S\times A\to\mathbb{R}} (1-\gamma)\,\mathbb{E}_{\mu_0\pi}\left[Q\left(s_0,a_0\right)\right]$$
$$\text{s.t.}\quad Q\left(s,a\right) \geqslant R\left(s,a\right) + \gamma\cdot\mathcal{P}^{\pi}Q\left(s,a\right),\quad \forall\left(s,a\right)\in S\times A,$$

Dual

$$\max_{d:S\times A\to\mathbb{R}_+} \mathbb{E}_d\left[r\left(s,a\right)\right]$$
$$\text{s.t.}\quad d\left(s,a\right) = (1-\gamma)\,\mu_0\pi\left(s,a\right) + \gamma\cdot\mathcal{P}_*^{\pi}d\left(s,a\right),\quad \forall\left(s,a\right)\in S\times A,$$

where the operator $\mathcal{P}^{\pi}$ and its adjoint, $\mathcal{P}_*^{\pi}$, are defined as

$$\mathcal{P}^{\pi}Q\left(s,a\right) := \mathbb{E}_{s'\sim T(\cdot|s,a),a'\sim\pi(\cdot|s')}\left[Q\left(s',a'\right)\right],$$
$$\mathcal{P}_*^{\pi}d\left(s,a\right) := \pi\left(a|s\right)\sum_{\widetilde{s},\widetilde{a}} T\left(s|\widetilde{s},\widetilde{a}\right)d\left(\widetilde{s},\widetilde{a}\right).$$

DICE Backbone

Lagrangian

$$\rho_\pi = \max_{\tau \geqslant 0} \min_{\nu} \ \mathbb{E}_{\mu_0 \pi, d^{\mathcal{D}}} \left[\ell \left(x; \tau, \nu \right) \right]$$

where $\tau \left(s, a \right) := \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$ is the *stationary DIstribution Corrector Estimation* and

$$\ell \left(x; \tau, \nu \right) := \tau(s, a) \cdot r(s, a) + (1 - \gamma) \, \nu \left(s_0, a_0 \right) + \tau \left(s, a \right) \left(\gamma \nu \left(s', a' \right) - \nu \left(s, a \right) \right)$$

DICE Backbone

Lagrangian

$$\rho_\pi = \max_{\tau \geqslant 0} \min_{\nu} \ \mathbb{E}_{\mu_0 \pi, d^{\mathcal{D}}} \left[\ell\left(x; \tau, \nu\right)\right]$$

where $\tau\left(s,a\right) := \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$ is the *stationary DIstribution Corrector Estimation* and

$$\ell\left(x;\tau,\nu\right) := \tau(s,a)\cdot r(s,a) + \left(1-\gamma\right)\nu\left(s_0,a_0\right) + \tau\left(s,a\right)\left(\gamma\nu\left(s',a'\right) - \nu\left(s,a\right)\right)$$

DICE family

The existing DICE family algorithms, *e.g.*, [NCDL19, ZDLS20, UHJ20, ZLW20], are the variants based on this Lagrangian [YND$^+$20].

Table of Content

Uncertainty is important

Optimism in the face of uncertainty [LS20]

Optimism in the face of uncertainty leads to *risk-seeking* algorithms, which can be used to balance the exploration/exploitation trade-off.

Pessimism in the face of uncertainty [SJ15, BGB20]

In offline reinforcement learning, a safe optimization criterion is to maximize the worst-case performance among a set of statistically plausible models

CoinDICE

Intuition from Bootstrap

- Contruct $\mathcal{D}_i$ by resampling from $\mathcal{D}$
- Run DICE estimator on $\mathcal{D}_i$, obtaining $\widehat{\rho}_i(\pi)$
- Estimate the variance from the set of estimators $\{\widehat{\rho}_i(\pi)\}_{i=1}^m$

CoinDICE

Intuition from Bootstrap

- Contruct $\mathcal{D}_i$ by resampling from $\mathcal{D}$
- Run DICE estimator on $\mathcal{D}_i$, obtaining $\widehat{\rho}_i(\pi)$
- Estimate the variance from the set of estimators $\{\widehat{\rho}_i(\pi)\}_{i=1}^m$

This procedure is computational expensive!

CoinDICE

Intuition from Bootstrap

- Contruct $\mathcal{D}_i$ by resampling from $\mathcal{D}$
- Run DICE estimator on $\mathcal{D}_i$, obtaining $\widehat{\rho}_i\left(\pi\right)$
- Estimate the variance from the set of estimators $\left\{\widehat{\rho}_i\left(\pi\right)\right\}_{i=1}^m$

This procedure is computational expensive!

Any way to reduce the compuation?

CoinDICE

Intuition from Bootstrap

- Contruct $\mathcal{D}_i$ by resampling from $\mathcal{D}$
- Run DICE estimator on $\mathcal{D}_i$, obtaining $\widehat{\rho}_i(\pi)$
- Estimate the variance from the set of estimators $\{\widehat{\rho}_i(\pi)\}_{i=1}^m$

This procedure is computational expensive!

Any way to reduce the compuation? YES!lol

CoinDICE (cont'd)

Optimizing the perturbation

$$[l_n, u_n] = \left[\min_{\nu} \max_{\tau \geqslant 0} \min_{w \in \mathcal{K}_f} \mathbb{E}_w \left[\ell\left(x; \tau, \nu\right)\right], \quad \max_{\tau \geqslant 0} \min_{\nu} \max_{w \in \mathcal{K}_f} \mathbb{E}_w \left[\ell\left(x; \tau, \nu\right)\right]\right]$$

$$\mathcal{K}_f := \left\{ w \in \mathcal{P}^{n-1}\left(\widehat{p}_n\right), \quad D_f\left(w || \widehat{p}_n\right) \leqslant \frac{\xi}{n} \right\} \tag{1}$$

CoinDICE (cont'd)

Optimizing the perturbation

$$[l_n, u_n] = \left[\min_{\nu}\max_{\tau\geqslant 0}\min_{w\in\mathcal{K}_f}\mathbb{E}_w\left[\ell\left(x;\tau,\nu\right)\right], \quad \max_{\tau\geqslant 0}\min_{\nu}\max_{w\in\mathcal{K}_f}\mathbb{E}_w\left[\ell\left(x;\tau,\nu\right)\right]\right]$$

$$\mathcal{K}_f := \left\{w\in\mathcal{P}^{n-1}\left(\widehat{p}_n\right), \quad D_f\left(w||\widehat{p}_n\right)\leqslant\frac{\xi}{n}\right\} \tag{1}$$

Closed-form reweighting

$$w_l = f'_*\left(\frac{\eta-\ell\left(x;\tau,\beta\right)}{\lambda}\right) \quad \text{and} \quad w_u = f'_*\left(\frac{\ell\left(x;\tau,\beta\right)-\eta}{\lambda}\right). \tag{2}$$

Connection to CVaR: With a special f selected, we recover the CVaR from the lower bound.

Theoretical Analysis

Asymptotic Coverage

Under mild conditions,

$$\lim_{n\to\infty} \mathbb{P}\left(\rho_\pi \in [l_n, u_n]\right) = \mathbb{P}\left(\chi^2_{(1)} \leqslant \xi\right). \tag{3}$$

Thus, $C^f_{n,\chi^{2,1-\alpha}_{(1)}} = [l_n, u_n]$ is an asymptotic $(1-\alpha)$-confidence interval of the value of the policy π.

Finite-sample Analysis

With high probability, we have

$$\rho_\pi \in \left[l_n - \mathcal{O}\left(\frac{1}{n}\right), u_n + \mathcal{O}\left(\frac{1}{n}\right)\right]. \tag{4}$$

Implementation of OFU and PFU

CoinDICE for OFU/PFU

- Estimate $(\beta_u^*, \tau_u^*, w_u^*)$ via CoinDICE for optimism. $//(\beta_l^*, \tau_l^*, w_l^*)$ for pessimism.
- Estimate the stochastic approximation to $\nabla_\pi u_{\mathcal{D}_t}(\pi_t)$. $//\nabla_\pi l_{\mathcal{D}_t}(\pi_t)$ for pessimism.
- Natural policy gradient update:
  $\pi_{t+1} = \mathrm{argmin}_\pi - \langle \pi, \nabla_\pi u_{\mathcal{D}_t}(\pi_t) \rangle + \frac{1}{\eta} KL(\pi || \pi_t)$.
  $//\pi_{t+1} = \mathrm{argmin}_\pi - \langle \pi, \nabla_\pi l_{\mathcal{D}_t}(\pi_t) \rangle + \frac{1}{\eta} KL(\pi || \pi_t)$ for pessimism.
- Collect samples $\mathcal{E} = \{x^{(j)} = (s_0, s, a, r, s')^{(j)}\}_{j=1}^m$ by executing π_{t+1}, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{E}$.
  //Skip the data collection step in offline setting.

Implementation of OFU and PFU

CoinDICE for OFU/PFU

- Estimate $(\beta_u^*, \tau_u^*, w_u^*)$ via CoinDICE for optimism. $//(\beta_l^*, \tau_l^*, w_l^*)$ for pessimism.
- Estimate the stochastic approximation to $\nabla_\pi u_{\mathcal{D}_t}(\pi_t)$. $//\nabla_\pi l_{\mathcal{D}_t}(\pi_t)$ for pessimism.
- Natural policy gradient update:
  $\pi_{t+1} = \mathrm{argmin}_\pi - \langle \pi, \nabla_\pi u_{\mathcal{D}_t}(\pi_t) \rangle + \frac{1}{\eta} KL(\pi || \pi_t)$.
  $//\pi_{t+1} = \mathrm{argmin}_\pi - \langle \pi, \nabla_\pi l_{\mathcal{D}_t}(\pi_t) \rangle + \frac{1}{\eta} KL(\pi || \pi_t)$ for pessimism.
- Collect samples $\mathcal{E} = \{x^{(j)} = (s_0, s, a, r, s')^{(j)}\}_{j=1}^m$ by executing π_{t+1}, $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \mathcal{E}$.
  //Skip the data collection step in offline setting.

Connection to Experience Replay: with different reweighting scheme, the expeience replay is for exploration or safe RL.

Experient Result

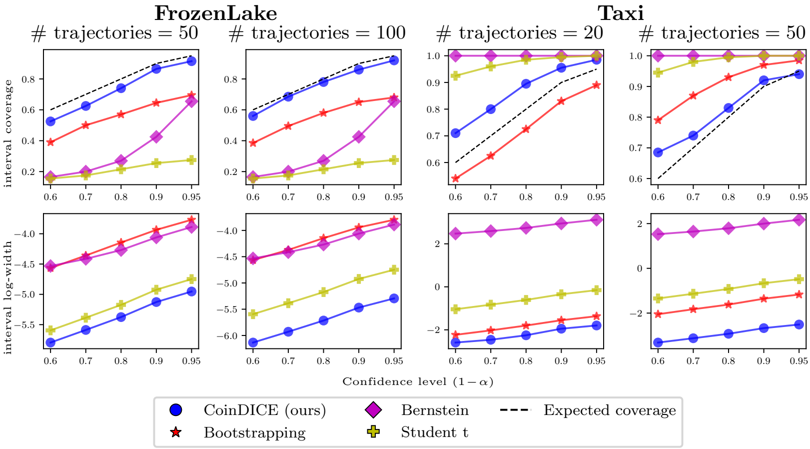


Table of Content

Conclusion

Recap

- We proposed a series of estimators for **behavior-agnostic** confidence interval estimation.
- These estimators can be used for implementing OFU/PFU.

Future work

- Regret bound of the OFU with CoinDICE (will release soon!)

Conclusion

Recap

- We proposed a series of estimators for **behavior-agnostic** confidence interval estimation.
- These estimators can be used for implementing OFU/PFU.

Future work

- Regret bound of the OFU with CoinDICE (will release soon!)
- Better algorithm for solving DICE.

Thanks!

[BGB20] Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[NCDL19] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. pages 2315–2325, 2019.

[SJ15] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[UHJ20] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. 2020.

[YND+20] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

[ZDLS20] Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. In *International Conference on Learning Representations*, 2020.

[ZLW20] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary value. *arXiv preprint arXiv:2001.11113*, 2020.