Evaluating Off-Policy Evaluation: Sensitivity and Robustness

Yuta Saito[1], Takuma Udagawa[2], Haruka Kiyohara[3], Kazuki Mogi[4], Yusuke Narita[5], Kei Tateno[2]

Hanjuku-Kaso Co., Ltd.[1], Sony Group Corporation[2], Tokyo Institute of Technology[3], Stanford University[4], Yale University[5]

Contact: Haruka Kiyohara
Tokyo Institute of Technology
Website: https://sites.google.com/view/harukakiyohara
Email: `kiyohara.h.aa@m.titech.ac.jp`

Bayesian Causal Inference for Real World Interactive Systems @ KDD 2021

Take-Home Message

- In applications such as recommender systems, we often want to **evaluate the performance of a policy in an offline manner** (OPE), without any risky online interaction.
- When applying OPE to a real-world problem, **we need to identify a robust estimator that works without significant hyperparameter tuning**.
- Identifying a robust estimator is extremely difficult with a typical experimental procedure used in OPE research.
- We develop a novel evaluation protocol, ***Interpretable Evaluation for Offline Evaluation***, which can **provide insights on the estimators' robustness**. (We also publicized a Python package, *pyIEOE*.)
- We **apply our procedure in a real-world e-commerce platform** and provide a suitable estimator choice for the platform.

Off-Policy Evaluation

We consider a general contextual bandit setting.

- $x \in \mathcal{X}$ is a context vector (e.g., the user's demographic profile)
- $a \in \mathcal{A}$ is an action (e.g., an item recommended from a finite set of items)
- $r \in [0, r_{\max}]$ is a reward (e.g., click indicator on the recommended item)

Decision making systems (e.g., recommender systems) are often constructed by a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, which chooses an action for each given context to maximize the following policy value (i.e., expected reward).

$$V(\pi) := \mathbb{E}_{(x,a,r)\sim p(x)\pi(a|x)p(r|x,a)}[r], \tag{1}$$

where $p(x)$ and $p(r \mid x, a)$ are unknown provability distributions.

Here, we assume that we have a historical logged bandit data obtained by a *behavior* policy π_b: $\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x)\pi_b(a \mid x)p(r \mid x, a)$, where n is the data size.

off-policy evaluation (OPE) **aims to evaluate the performance of a counterfactual *evaluation* policy** π_e using only $\mathcal{D}$ as follows.

$$\hat{V}(\pi_e; \mathcal{D}, \theta) \approx V(\pi_e), \tag{2}$$

where $\hat{V}$ is an OPE estimator, and θ is a set of estimator's (pre-defined) hyperparameters. Below, we show several examples of OPE estimators.

- **IPW** mitigates distribution shift between π_b and π_e using importance sampling techniques as $\hat{V}_{\mathrm{IPW}} := \mathbb{E}_n[\rho(x_i, a_i)r_i]$, where $\mathbb{E}_n[\cdot]$ is empirical average over $\mathcal{D}$ and $\rho(x_i, a_i) := \pi_e(x_i \mid a_i)/\pi_b(x_i, a_i)$ is the importance weight. This estimator is hyperparameter free but can suffer from large variance.

Algorithm 1 Interpretable Evaluation for Offline Evaluation

Input: logged bandit feedback $\mathcal{D}$, an estimator to be evaluated $\hat{V}$, a candidate set of hyperparameters Θ, a set of evaluation policies Π_e, a hyperparameter sampler ϕ (default: uniform distribution), a set of random seeds $\mathcal{S}$
Output: empirical CDF, $\hat{F}_Z$, of the squared error (SE)
$\mathcal{Z} \leftarrow \emptyset$
for $s \in \mathcal{S}$ **do**
  $\theta \leftarrow \phi(\Theta; s)$
  $\pi_e \leftarrow \mathrm{Unif}(\Pi_e; s)$
  $\mathcal{D}^* \leftarrow \mathrm{Bootstrap}(\mathcal{D}; s)$
  $z' \leftarrow \mathrm{SE}(\hat{V}; \mathcal{D}^*, \pi_e, \theta)$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{z'\}$
end for
Estimate F_Z using $\mathcal{Z}$ (by Eq. 1)

- **SNIPW** tries to address the variance of IPW by dividing $\hat{V}_{\mathrm{IPW}}$ by the sum of importance weights as $\hat{V}_{\mathrm{SNIPW}} := \mathbb{E}_n[\rho(x_i, a_i)r_i]/\mathbb{E}_n[\rho(x_i, a_i)]$. This estimator is also hyperparameter free.
- **DR** also attempts to tackle the variance of IPW by leveraging baseline estimation $\hat{q}$ and perform importance weighting only on its residual as $\hat{V}_{\mathrm{DR}} := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x_i)}[\hat{q}(x_i, a)] + \rho(x_i, a_i)(r_i - \hat{q}(x_i, a_i))]$. To use DR, we have to set hyperparameters of $\hat{q}$.
- **Switch-DR** aims to further reduce variance of DR by avoiding importance weighting when ρ is large as $\hat{V}_{\mathrm{Switch-DR}} := \mathbb{E}_n[\mathbb{E}_{a\sim\pi_e(a|x)}[\hat{q}(x_i, a)] + \rho(x_i, a_i)\mathbb{I}\{\rho(x_i, a_i) \le \tau\}(r_i - \hat{q}(x_i, a_i))]$. This estimator have two hyperparameters, $\hat{q}$ and τ.

Conventional Evaluation and Limitation

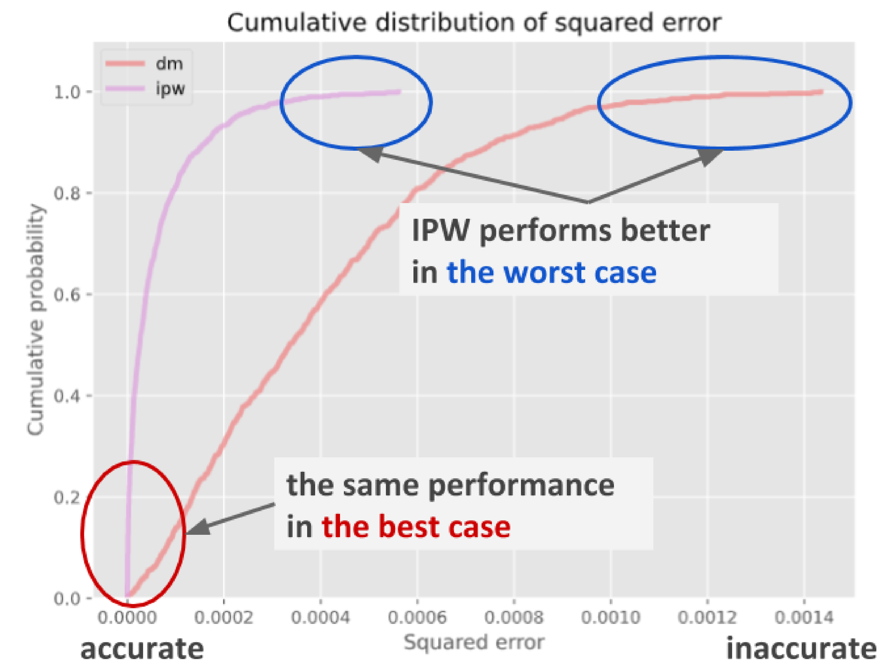
To evaluate and compare the performance of OPE estimators, we use the following *squared-error* (SE) as a performance measure for any given $\pi_e, \mathcal{D}, \hat{V}, \theta$.

$$\mathrm{SE}(\hat{V}; \pi_e, \theta) := (V(\pi_e) - \hat{V}(\pi_e; \mathcal{D}, \theta))^2, \tag{3}$$

A typical evaluation procedure calculates *mean-squared-error* (MSE) for a single given set of $(\pi_e, \mathcal{D}, \theta)$ to compare the performance of OPE estimators. We argue that **this typical procedure cannot evaluate the estimators' robustness to the configuration changes** of $(\pi_e, \mathcal{D}, \theta)$.

Interpretable Evaluation for Offline Evaluation (IEOE)

To measure the estimators' robustness, we first calculate SE on a various set of configurations as shown in Algorithm 1. Moreover, we use ***cumulative distribution function* (CDF) to conduct a more informative comparison** of the estimators' performance. CDF is a function defined as $F_Z(z) := \mathbb{P}(Z \le z)$, which is the probability that the estimator achieves a performance better or equal to z. When we have $\mathcal{Z} = \{z_1, \ldots, z_m\}$ (which corresponds to SE), we can estimate CDF as follows.

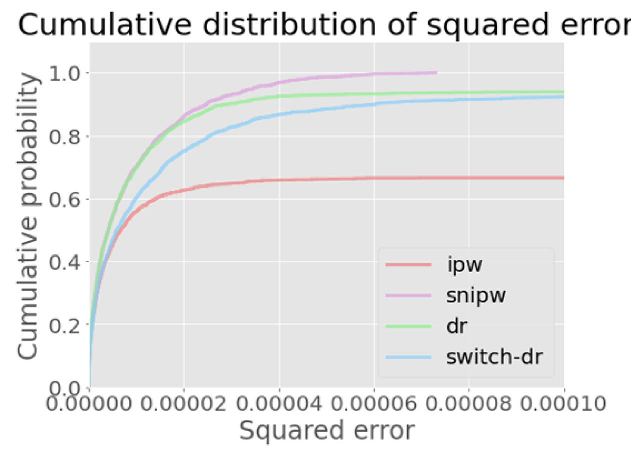


$$\hat{F}_Z(z) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\{z_i \le z\}, \tag{4}$$

We can **visualize CDF for an interpretable comparison** as we show in the above figure. Using CDF, we also define evaluation metrics such as *area under the CDF curve* $\text{AU-CDF}(z_{\max}) := \int_0^{z_{\max}} F_Z(z)dz$ and *conditional value-at-risk* $\mathrm{CVaR}_\alpha(Z) := \mathbb{E}[Z \mid Z \ge F_Z^{-1}(\alpha)]$, which will be useful for identifying the robust OPE estimators.

Real World Application

We applied the IEOE procedure to provide a suitable estimator choice for a real e-commerce platform. In the experiment, we found SNIPW clearly outperforms other estimators across various configurations on the platform data. **The platform is now using SNIPW after the comprehensive accuracy and stability verification with IEOE**.



OPE Estimators	AU-CDF	$\mathrm{CVaR}_{0.7}$
IPW	**0.721**†	**32.86**†
SNIPW	**1.000***	**1.00***
DR	0.972	11.25
Switch-DR	0.882	12.26

Note: Larger value of AU-CDF and lower value of CVaR indicate that the estimator is more accurate. We use $z_{max} = 5.0 \times 10^{-5}$ and $\alpha = 0.7$. The colors correspond to **best** and **worst**. The value is divided by that of the best estimator.

Check out our camera-ready/arXiv paper for more detailed results! Also, feel free to ask any questions.